École Nationale Supérieure des Sciences de
l'Information et des Bibliothèques

ENSSIB

DEA

Sciences de l'Information et de la Communication

option : systèmes d'information documentaire

MÉMOIRE DE DEA

Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux

VOLUME I

Réalisé par Hélio KURAMOTO
boursier du CNPq, Brasília - Brésil

sous la direction de M. Michel LE GUERN
Professeur à l'Université Lumière - Lyon 2

Septembre / 1995

Université Lumière Lyon 2 | École Nationale Supérieure des Sciences de l'Information et des Bibliothèques | Université Jean-Moulin Lyon 3

École Nationale Supérieure des Sciences de
l'Information et des Bibliothèques

ENSSIB

DEA

Sciences de l'Information et de la Communication

option : systèmes d'information documentaire

MÉMOIRE DE DEA

Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux

VOLUME I

Réalisé par Hélio KURAMOTO

boursier du CNPq, Brasília - Brésil

sous la direction de M. Michel LE GUERN

Professeur à l'Université Lumière - Lyon 2

Septembre / 1995

Université Lumière Lyon 2 | École Nationale Supérieure des Sciences de l'Information et des Bibliothèques | Université Jean-Moulin Lyon 3

Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux

Hélio KURAMOTO

sous la direction de M. Michel LE GUERN

Résumé : Ce travail porte sur une approche alternative de traitement et de recherche d'information où les syntagmes nominaux jouent le rôle principal. Il s'agit d'une maquette de système de recherche interactif d'information capable de naviguer dans une structure arborescente de syntagmes nominaux. Ceux-ci ont été extraits manuellement d'un corpus en langue portugaise, en utilisant un raisonnement logico-sémantique.

Descripteurs français : Syntagmes Nominaux ; Indexation ; Systèmes de Recherche d'Information ; Modèle Relationnel de données.

Abstract : This work provide an alternative approach of processing and retrieving information, in which the main role is played by the nominal syntagms. An interactive system of retrieving information was developed to be able to navigate into the nominal syntagms structures. These syntagms are manually drawn from a corpus in portuguese language using a logic-semantic procedure.

English Keyword : Nominals syntagms ; Indexing ; Information Retrieval Systems ; Relational Data Model.

Je dédie ce travail à
mes femmes Juçara, Liliana et Luisa
qui m'ont beaucoup aidé avec leur
patience, confiance et tendresse.

Je remercie

M. Michel LE GUERN

pour ses enseignements, pour sa sagesse,
pour sa patience, pour m'avoir confié ce travail
et pour me faire partager son expérience
dans la direction de ce mémoire.

M. Richard BOUCHÉ

pour m'avoir accepté dans le cours de DEA,
pour les enseignements, pour le support
et pour l'orientation de ma Note de Synthèse.

Mme. Maria Eugênia Malheiros POULET

pour m'orienter sur questions de langue portugaise.

M. José Rincon FERREIRA

pour me donner l'opportunité d'étudier en France.

Mme. Maria Lina Pereira e SILVA

pour son amitié, son support moral
et dans le cadre de la langue française
en révisant ce travail.

Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq

pour m'avait accorder une bourse me permettant
d'entreprendre ces études.

M. Paulo César SIQUEIRA

pour son amitié, son support moral
et pour sa contribution en révisant le résumé en anglais.

Table de Matières

Volume I

Volume II

Volume III

Table des illustrations

Avant Propos

Pendant les dernières années, des milliards d'informations ont été enregistrées dans plusieurs bases de données, dans les domaines les plus divers de connaissances et sur diverses formes (numériques, textuelles, imagées etc...). Etant donné que les ressources informationnelles sont de plus en plus accessibles à l'usage privé, le principal problème qui se pose aujourd'hui est de savoir comment accéder à l'information dont on a besoin, de façon précise et conviviale. Pour cela, il faut donc utiliser des systèmes de recherche d'information.

Les usagers ont du mal à utiliser de tels systèmes, soit parce qu'ils ne sont toujours pas faciles ou conviviales à opérer, soit à cause des résultats qui ne sont pas toujours précis. En ce qui concerne l'utilisation de ces systèmes, les difficultés sont souvent liées aux langages de recherche et aux complexités de leurs expressions de recherche d'information. D'autre part, les résultats peu précis sont en quelque sorte conséquence du processus de traitement de l'information.

Il faut donc trouver de nouvelles approches pour améliorer ces systèmes en considérant les deux aspects présentés ci-dessus. Le travail que l'on propose concerne l'expérimentation d'une approche alternative de traitement et de recherche d'information où les syntagmes nominaux jouent le principal rôle dans la procédure de navigation et d'interrogation d'une base de données texte plein. Les principaux buts à accomplir sont : a) acquérir des connaissances sur l'extraction des syntagmes nominaux dans des documents en langue portugaise, envisageant d'ores et déjà le développement futur d'un analyseur

morpho-syntaxique pour l'extraction automatique des syntagmes nominaux sur des documents écrits en cette langue ; b) dessiner et développer une maquette de système de recherche d'information capable d'utiliser les syntagmes nominaux comme un moyen de recherche d'information ; c) exploiter ce système de façon à connaître le comportement des syntagmes nominaux dans la procédure d'interrogation d'une base de données texte plein.

Introduction

La technologie de développement d'un système de recherche d'information textuel est aujourd'hui bien connue. D'une façon générale elle comprend les processus de traitement de l'information, de stockage et d'interaction avec les utilisateurs. Strzalkowski décrit une tel système comme :

> *« Un Système de Recherche d'Information typique sélectionne les documents d'une base de données en réponse aux requêtes des utilisateurs et les présente selon un ordre de pertinence. La procédure habituelle consiste d'une part d'une méthode d'indexation afin de sélectionner les termes (mots ou phrases), et d'autre part en la création d'un fichier inversé d'indices par lequel l'accès facile aux documents sera fait. »* [1]

Cette description, bien que très simple et résumé, montre que derrière un tel système il y a des outils de représentation du contenu d'une base de données, de sa structure et du stockage de ces contenus. Alan Smeaton, résume à son tour la représentation du contenu, comme suit :

> *« L'approche conventionnelle pour la représentation du contenu des documents est réalisée soit par l'indexation automatique du texte par mots ou racines de mots soit par l'indexation par mots ou phrases extraits manuellement d'un vocabulaire contrôlé. »*[2]

Pour qu'un système de recherche d'information puisse répondre aux demandes des utilisateurs dans des délais acceptables, il faut que ces documents soient soumis à un traitement préalable. Cela permet la structuration des informations de manière à accéder aux

---

[1] Tomek STRZALKOWSKI. « Natural language processing in large-scale text retrieval tasks ». *Text Retrieval Conference (TREC-1)*. Gaithersburg, 1993. p. 173-187.

[2] Alan F. SMEATON. « Prospects for intelligent, languaged-based information retrieval ». *Online Review*. 1991, vol. 15, n°. 6. p. 373-382.

documents d'une forme précise et rapide. On a vu que le traitement préalable de l'information consiste en l'indexation automatique et/ou manuel. Les produits de cette procédure sont les mots et les racines de mots. On ne va traiter ici que de l'indexation automatique, puisqu'elle s'applique plus aux bases de données texte plein, objet de ce travail.

L'indexation automatique peut se faire selon plusieurs méthodes. On peut distinguer les suivantes :

a) « *des méthodes élémentaires comme la simple extraction des mots où ces mots sont des descripteurs en excluant les mots vides* »[3]

b) « *[des méthodes statistiques que] consistent à définir un modèle probabiliste des occurrences des mots dans un document considéré à l'intérieur d'une collection bien définie de manière à établir leur caractère pertinent pour participer à la description du contenu* »[4]

c) « *des méthodes linguistiques où l'identification des parties du discours pertinentes se fait sur des critères linguistiques... [Une des approches linguistiques part du principe que] le lexique, en tant que composant de la langue, ne contient que des éléments qui sont des propriétés, c'est-à-dire des prédicats. Le mot est donc un prédicat et il ne peut pas, considéré isolément, faire référence à un objet de la réalité extra-linguistique de l'auteur du document. Il ne peut pas exprimer "ce dont parle le documents". Il ne peut donc pas être un descripteur* »[5].

> « *Comme dit Michel Le Guern " La finalité du descripteur exclut qu'on puisse l'envisager en faisant abstraction de la valeur référentielle de ses occurrences dans le corpus. Les mots de la langue, en tant qu'ils sont*

---

[3] Christian FLUHR. « Le traitement du langage naturel dans la recherche d'information documentaire ». In.: *Cours INRIA - Interfaces Intelligentes dans l'Information Scientifique et Technique*. 18-22 Mai 1992. p. 103-128

[4] Richard BOUCHÉ. « Le Syntagme Nominal, une Nouvelle Approche des Bases de Données Textuelles ». *Meta*. 1989, vol. 34, n°. 3. p. 428.

[5] Idem p. 429.

mots de la langue, ne signifient que des attributs, et non des substances, tant qu'ils ne sont pas mis en oeuvre dans le discours. Le descripteur, quant à lui signifie une entité au sens de la philosophie d'Aristote. Le descripteur ne peut donc pas être considéré, à l'instar des mots de la langue comme un symbole sans référence." [6]

« On admettra donc que la plus petite unité du discours porteuse d'une valeur référentielle est le syntagme nominal. C'est elle qu'importe d'identifier dans le document. »[7]

Les approches « a » et « b », plus haut, sont les plus communément utilisées pour les grandes banques de données textuelles. Étant donné l'objet de ce travail, on ne va discuter ici que l'approche « c ».

L'utilisation des syntagmes nominaux comme descripteurs possède quelques avantages de même qu'un potentiel d'organisation par rapport à ceux obtenus à partir des méthodes « a » ou « b ». Les syntagmes nominaux peuvent être organisés de manière à rendre plus facile, aux utilisateurs, la tâche d'effectuer la recherche d'information. M. Le Guern le démontre dans un article publié dans la revue *Le Français Moderne,* juin 1991 :

« ... Et on peut se demander à quoi servent ces syntagmes nominaux une fois qu'ils ont été extraits. En soi, la liste de tous les syntagmes nominaux de corpus, accompagnés pour chacun de la liste des références de ses occurrences, est déjà utile. Mais, pour une plus grande efficacité de l'outil d'interrogation, il convient de structurer cet ensemble de syntagmes nominaux, en se servant d'abord des relations d'emboîtement...

« ... L'outil d'aide à l'interrogation donne, pour chaque syntagme d'un niveau donné, la liste des syntagmes de rang immédiatement supérieur qui le contiennent, avec le nombre d'occurrences de chaque syntagme dans le corpus. L'utilisateur peut ainsi cheminer dans l'ensemble des descripteurs, en réduisant progressivement le nombre d'occurrences,

---

[6] Michel LE GUERN. « Les descripteurs d'un système documentaire : essai de définition », In. : Bès, G.C., Fauchère, P.M., Lagueunière, F. *Actes du Colloque « Traitement automatique des langues naturelles et systèmes documentaires ».* Condenser, supplément I,Université Clermont Ferrand, 1982.

[7] Richard BOUCHÉ. « Le Syntagme Nominal, une Nouvelle Approche des Bases de Données Textuelles ». *Meta.* 1989, vol. 34, n°. 3. p. 430.

jusqu'au moment où il obtiendra la quantité de références correspondant à l'ordre de grandeur qu'il souhaite. »[8]

Ainsi les syntagmes nominaux peuvent être utilisés dans un système de recherche d'information différemment des descripteurs. Ils peuvent être organisés dans une structure de manière à ce que les usagers puissent y naviguer. La proposition de Le Guern est claire, simple et souple. L'utilisateur est amené à trouver les informations d'une forme simple, convergente et précise. Le Guern le montre en utilisant un exemple extrait du journal *Le Monde*, du vendredi 30 mars 1990, page 31, colonne 2 :

Le rapport Gilles Bélier sur « les conditions de l'amélioration de la représentation des salariés dans les PME ».

L'analyse fournit 6 syntagmes repartis en 5 niveaux :

(5) *Le rapport Gilles Bélier sur « les conditions de l'amélioration de la représentation des salariés dans les PME »*

(4) *les conditions de l'amélioration de la représentation des salariés dans les PME*

(3) *l'amélioration de la représentation des salariés dans les PME*

(2) *la représentation des salariés dans les PME*

(1) *les salariés*

(1) *les PME*

Sur cet exemple, Le Guern considère que :

---

[8] Michel LE GUERN. « Un analyseur morpho-syntaxique pour l'indexation automatique », *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1, p. 34

«... Le cheminement dans l'ensemble des descripteurs, à partir de l'entrée "Les PME", permettra à l'utilisateur de retrouver l'indication sur "le rapport Bélier", à partir de laquelle il pourra lancer une interrogation complémentaire.

« On pourra objecter que ces cheminements sont inutilement fastidieux, et que l'utilisateur pourrait se contenter de demander quelles sont dans le corpus les occurrences du syntagme "la représentation des salariés dans les PME". L'ennui, c'est qu'il manquerait les occurrences du syntagme "la représentation des salariés au sein des PME", qui figure lui aussi dans le corpus. Or, à partir des syntagmes de niveau (1) "les salariés" et "les PME", le cheminement fournit aussi bien la forme avec "au sein de" que la forme avec "dans les". Supposons maintenant que l'utilisateur demande "l'accroissement de la représentation des salariés dans les PME", et que ce syntagme ne figure pas littéralement dans le corpus ; il n'aura aucune réponse. La liste des syntagmes du rang supérieur à "la représentation des salariés dans les PME" lui propose "l'amélioration de la représentation des salariés dans les PME", ce qui le satisfait. Ainsi se trouve contourné dans de nombreux cas l'obstacle majeur que constitue la parasynonymie pour les systèmes d'information textuelle. »[9]

Prenant en compte l'approche proposé par Le Guern, on a développé une maquette d'un système de recherche d'information utilisant des syntagmes nominaux. Pour que l'on puisse accomplir et expérimenter une telle maquette il a été nécessaire de construire une base de données. Ainsi, ce travail comprend les trois phases suivantes : 1) construction de la base de données avec un corpus en langue portugaise ; 2) développement de la maquette du système de recherche d'information ; 3) exploitation du système de recherche d'information.

Dans la phase 1 on a définit les critères pour la construction de la base de données, le traitement préalable du corpus et l'extraction des syntagmes nominaux. La phase 2 englobe les tâches du choix de l'approche pour la construction de la maquette, ainsi que sa construction effective. La dernière phase a été dédiée à l'exploitation de la maquette du système de recherche d'information, au moyen du chargement de la base de données, la

[9] Michel LE GUERN. « Un analyseur morpho-syntaxique pour l'indexation automatique », *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1, p. 34-35.

construction de l'arborescence, l'établissement des centres des syntagmes nominaux pour chaque syntagme de premier niveau et la vérification de l'interface de recherche d'information. Les phases de construction et d'exploitation de la maquette ont été développées en deux étapes. Dans la première, une fois la maquette construite, on l'a testée au moyen d'un échantillon du corpus (cinq articles). Il va sans dire que cette étape a été très importante pour connaître les vrais limites du logiciel choisi de même que le comportement des syntagmes nominaux dans l'arborescence. Les résultats de ces observations ont été utilisés dans la deuxième étape pour l'ajustement nécessaire de la maquette définitive.

« Une proposition incorrecte est forcément fausse, mais une proposition correcte n'est pas forcément vraie »

Emmanuel KANT (1724 -1804)

Chapitre Premier

Construction de la base de données

Avant de commencer la construction de la base de données et de la maquette proprement dite, des critères pour le choix du corpus ont été établis en tenant compte du temps disponible pour achever le travail, des conditions nécessaires pour le maîtriser et le réussir, et en considérant la durée du cours de DEA, c'est-à-dire environ six (6) mois à temps partiel et deux (2) mois à plein temps. Pour la construction de la base de données proposée il a fallu donc, comme on l'a déjà souligné, mettre en oeuvre les étapes suivantes : (a) établissement de critères pour le choix du corpus ; (b) traitement préalable des articles ; (c) extraction des syntagmes nominaux.

1. Etablissement de critères pour le choix du corpus

Les critères relatives aux conditions nécessaires pour maîtriser ce travail sont donc les suivants :

a) Taille du corpus

Pour une expérimentation de recherche d'information on a trouvé que quinze articles (voir Annexe A), constituaient le nombre minimal susceptible d'être traité dans la limite du temps disponible.

b) Taille des articles

Les articles choisis ont de trois à cinq pages, dans les fichiers format Word 6.0a, police Times New Roman, style normal, taille 11. Cependant on a eu des difficultés à les trouver dans cette taille, ce qui a amené à supprimer quelques paragraphes tout en gardant le soin de ne pas perdre les syntagmes nominaux importants.

c) Le domaine de connaissance du corpus

Un système de recherche d'information utilisant des syntagmes nominaux sera plus performant s'il travaille sur un domaine bien définit. Cette restriction est nécessaire car il faut travailler sur un corpus homogène par rapport à l'ensemble des syntagmes nominaux, avec un minimum d'ambiguïtés, de façon à ce qu'on puisse les organiser sous forme d'arborescence. Selon Minsky :

> *« Dans le langage naturel, les ambiguïtés ne découlent pas seulement du fait que les mots peuvent être regroupés de diverses façons, mais encore de ce que chaque mot peut avoir plusieurs sens différents... »*[10]

Ainsi, la définition du domaine du corpus d'une base de données est d'une importance capitale pour que les résultats de recherche soient précis. Pour les bases de données multidisciplinaires la bonne solution serait plutôt de les partager en plusieurs bases de données regroupés par domaines de connaissances.

On a donc ainsi choisi pour ce travail le domaine des Sciences de l'Information (en considérant ici la pluridisciplinarité de ce domaine).

---

[10] Marvin MINSKY. *Semantic Information Processing*. Cambridge, Mass. : M.I.T. Press, 1969, p. 18, cité par Hubert L. Dreyfus dans son ouvrage *Intelligence Artificielle : mythes et limites*

d) Langue du corpus

On a choisi la langue portugaise pour deux raisons :

- afin d'acquérir des connaissances sur l'extraction des syntagmes nominaux dans cette langue et envisageant d'ores et déjà, dans le cadre d'une thèse de doctorat, la possibilité de développement d'un analyseur morpho-syntaxique pour cette langue ;

- du fait que mon but initial a été de travailler sur le développement de systèmes de recherche d'information sur des bases de données en langue portugaise. C'est aussi un compromis personnel avec l'institution dont j'ai obtenu ma bourse d'études.

Cependant, les résultats de ce travail pourront servir à des corpus dans d'autres langues, puisque l'extraction des syntagmes nominaux a été faite manuellement et la maquette est indépendante du traitement de l'information.

e) Niveaux des syntagmes nominaux

Les syntagmes nominaux possèdent des relations d'emboîtement les uns par rapport aux autres. L'ordre de la relation d'emboîtement, appelé niveau, détermine la hauteur de l'arbre des syntagmes nominaux qui à son tour, restreint les possibilités de raffinement de la recherche d'information. Afin de construire l'arborescence permettant le raffinement d'une recherche d'information, on a choisi des articles ayant au moins des syntagmes nominaux de niveau quatre.

Une fois ces critères établis, on a choisi des articles publiés dans la revue brésilienne « Ciência da Informação », spécialisée en Sciences de l'Information. Cette revue est publiée et distribuée par l'Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Les articles sélectionnés ont comme titres :

1. Conhecimento como recurso estratégico empresarial[11] (La connaissance comme ressource stratégique des entreprises) - mots clés : ressources informationnelles ; intelligence compétitive ;

2. Inteligência competititva e decisão empresarial[12] (L'intelligence compétitive et la décision des entreprises) - mots clés : information ; intelligence compétitive ; gestion ; stratégies de décision ;

3. Economia da informação (L'économie de l'information)[13] - mots clés : économie de l'information ; information / caractéristiques ; analyse du coût-bénéfice / coût / efficacité / performance / valeur ;

4. A Informação como insumo estratégico[14] (L'information comme matière première stratégique) - mots clés : information stratégique ; systèmes d'information ; information opérationnelle ; gestion stratégique ;

5. Informação técnico-econômica: mais importante do que nunca[15] (L'information téchnique-économique : plus important que jamais) - mots clés : information technologique ; information économique ; systèmes d'information téchnique-économique ; politique de recherche et de développement / entreprises ;

6. Perspectivas do agente da informação no contexto brasileiro[16] (Perspectives de l'agent de l'information dans le contexte brésilien) - mots clés : agent de l'information ; bibliothécaire ; spécialiste de l'information ;

---

[11] Anna Soledade VIEIRA. « Conhecimento como recurso estratégico empresarial ». *Ciência da Informação*. 1993, vol. 22, nº 2. p. 99-101.

[12] Patrick MAURY. « Inteligência competitiva e decisão empresarial ». *Ciência da Informação*. 1993, vol. 22, nº 2. p. 138-141.

[13] Pedro Onofre FERNANDES. « Economia da Informação ». *Ciência da Informação*. 1991, vol. 20, nº 2. p. 165-168.

[14] Dorodame Moura LEITÃO. « A informação como insumo estratégico ». *Ciência da Informação*. 1993, vol. 22, nº 2. p. 118-123.

[15] João Salvador FURTADO. « Informação técnico-econômica : mais importante do que nunca ». *Ciência da Informação*. 1991, vol. 20, nº 1. p. 20-22.

[16] Denise Werneck de PAIVA. « Perspectivas do agente da informação no contexto brasileiro ». *Ciência da Informação*. 1990, vol. 19, nº 1. p. 48-52.

7. Sistemas de informação : a evolução dos enfoques[17] (Les systèmes d'information : l'évolution de ses approches) - mots clés : systèmes d'information ; théorie des systèmes ; services d'information ;

8. Consultoria informatológica em revisão : uma alternativa para serviços de informação personalizados[18] (Consultation dans le domaine des sciences de l'information en révision : une alternative pour les services d'information personnalisés) - mots clés : services d'information ; bibliothèques spécialisées ; consultation dans le domaine des sciences de l'information ;

9. Informação para a indústria[19] (L'information pour l'industrie) - mots clés : information industrielle ; transfert de l'information ; information technologique ; information technologique / petite et moyenne industrie / Brésil ;

10. Interação entre empresas com necessidades de informação (=conhecimento) e a estrutura nacional de centros com provisão de conhecimento acumulado : referência especial à estrutura nacional de serviços de informação, documentação e de biblioteca[20] (Interaction entre les entreprises ayant besoin d'information (=connaissances) et la structure nationale de centres ayant un fonds : référence spéciale à la structure de services d'information, de documentation et de bibliothèques) - mots clés : politique d'information ; transfert d'information ; flux d'information ; centres et services d'information ; information technologique ;

---

[17] Marcos DANTAS. « Sistemas de Informação : a evolução dos enfoques ». *Ciência da Informação*. 1992, vol. 21, nº 3. p. 192-196.

[18] Mariano A. MAURA. « Consultoria Informatológica em revisão : uma alternativa para serviços de informação personalizados ». *Ciência da Informação*. 1993, vol. 22, nº 3. p. 242-247.

[19] Marisa Gurjão PINHEIRO. « Informação para a Indústria ». *Ciência da Informação*. 1991, vol. 20, nº 1. p. 16-19.

[20] Kjeld KLINTOE. « Interação entre empresas com necessidades de informação (=conhecimento) e a estrutura nacional de centros com provisão de conhecimento acumulado : referência especial à estrutura nacional de serviços de informação, documentação e de biblioteca ». *Ciência da Informação*. 1991, vol. 20, nº 1. p. 55-57.

11. Uso da informação na indústria como paradigma para o desenvolvimento econômico[21] (L'utilisation de l'information dans l'industrie comme paradigme pour le développement) - mots clés : information / développement économique ; information technologique ; information industriel ; information économique ; services d'information / entreprise ;

12. A Informação eficaz na empresa[22] (L'information efficace dans l'entreprise) - mots clés : information technologique ; prospection technologique ; services d'information ; entreprise de consultation ; entraînement de gestion ; ressources humaines ; produits d'information ;

13. Gerência da informação: mudanças nos perfis profissionais[23] (La gestion de l'information : changement dans les profils professionnels) - mots clés : administration des ressources d'information ; gestion de l'information ; professionnel de l'information ;

14. Informação: instrumento de dominação e de submissão[24] (L'information : outil de domination et de soumission) - mots clés : transfert d'information ; information technologique ; politique d'information ; développement technologique ; politique de science et technologie ; transfert de technologie ;

15. Informação: a chave para a qualidade total[25] (L'information : la clé pour la qualité totale) - mots clés : qualité totale ; information pour la qualité ; unités d'information ; systèmes d'information.

---

[21] Francisco das Chagas de SOUZA. « Uso da informação na indústria como paradigma para o desenvolvimento econômico ». *Ciência da Informação.* 1991, vol. 20, nº 1. p. 34-36.

[22] Auta Rojas BARRETO. « A informação eficaz na empresa ». *Ciência da Informação.* 1991, vol. 20, nº 1. p. 78-81.

[23] Regina de Barros CIANCONI. « Gerência da informação : mudança nos perfis profissionais ». *Ciência da Informação.* 1991, vol. 20, nº 2. p. 204-208.

[24] Vânia Maria Rodrigues de ARAÚJO. « Informação: instrumento de dominação e de submissão ». *Ciência da Informação.* 1991, vol. 20, nº 1. p. 37-43.

[25] Virginia Bentes PINTO. « Informação : a chave para a qualidade total ». *Ciência da Informação.* 1993, vol. 22, nº 2. p. 133-137.

Remarque : les articles ayant été indexés dans la revue Ciência da Informação, les mots-clés ont été traduits.

2. Traitement préalable du corpus

2.1 Saisie des articles

Pour gagner du temps, dans la procédure d'extraction des syntagmes nominaux et du chargement de la base de données dans la maquette, il a fallu saisir les articles sur l'ordinateur et les préparer au préalable pour le traitement de l'information. On a voulu travailler sur l'ordinateur à partir du moment où les articles ont été choisis.

Pour la saisie des articles sur l'ordinateur, on a utilisé un *scanner* à main *(ScanMan Logitech)* pour la numérisation des textes et le logiciel Omnipage Direct pour la reconnaissance des caractères. Ce *scanner* a été choisi pour la bonne raison qu'il a été livré avec un logiciel d'OCR (*Optical Character Recognition*) - ou logiciel de reconnaissance optique des caractères - capable de reconnaître les caractères de la langue portugaise. En fait ce logiciel traite des textes en onze langues (allemand, anglais, danois, espagnol, français, italien, irlandais/gaélique, néerlandais, norvégien, portugais, suédois). L'autre raison déterminante de ce choix était le prix, moins cher que les *scanners* à plat. Ainsi, bien que l'on sache que des questions d'ergonomie et de précision se posent pour ce type de *scanner*, on l'a quand même choisi car il n'y avait que quinze articles à traiter.

Le temps nécessaire pour numériser chaque article a été d'environ trois heures en moyenne. Ce temps, un peu long, s'explique car la bonne utilisation de ce *scanner* dépende fondamentalement de la dextérité de la main de l'opérateur. En outre, la qualité de l'impression du document a être numérisé et le réglage du contraste du *scanner* comptent beaucoup pour la précision de la numérisation et de la reconnaissance des caractères d'un texte.

Les caractères non reconnus par le logiciel d'OCR ont été remplacés par le caractère « @ ». Cependant, d'autres caractères ont été reconnus d'une manière incorrecte. On a trouvé les problèmes suivants :

a) quelques lettres accentuées sont souvent prises pour d'autres lettres en fonction de la proximité de l'accent à la lettre. Exemple : la lettre « ó » a été parfois reconnue comme un « 6 », la lettre « í » comme la lettre « f » et quelquefois comme la lettre « r » ;

b) la lettre « r » proche de la lettre « n » a été parfois reconnue comme étant la lettre « m » ;

c) l'inverse de la situation 'b', plus haut, a été constaté lorsque le mot « information » apparaissait dans le texte et que l'OCR l'a reconnu comme étant « inforrnation ». Il semble reconnaître la lettre « m » comme étant les lettres « r » et « n » ;

d) la lettre « i » est parfois reconnue comme étant la lettre « l ».

Cette expérience a montré que pour un travail professionnel, il faut plutôt utiliser un *scanner* à plat et un logiciel d'OCR capable de résoudre les problèmes orthographiques dus à la méconnaissance des caractères, car la correction automatique de l'orthographe peut conduire le logiciel à adopter des mots qui n'ont rien à voir avec les mots du texte. Il faut donc choisir un logiciel qui puisse proposer aux utilisateurs le choix du mot correct, c'est-à-dire, un logiciel avec un minimum d'interactivité avec l'utilisateur ou l'opérateur du *scanner*.

2.2 Préparation des fichiers des articles

Pour le traitement des articles, on a choisi les logiciel Word version 6.0a. La préparation des articles a consisté à énumérer chaque paragraphe dans chaque article du corpus. Cette procédure a été nécessaire afin d'identifier chaque syntagme nominal; cette

identification des syntagmes nominaux permettra à son tour de retrouver les articles d'où ils ont été extraits. Ainsi, les syntagmes nominaux ont été identifiés par le numéro d'article et le numéro d'ordre du paragraphe d'où ils ont été extraits.

Exemple : ***a informação; 1; 4***

qui veut dire que le syntagme ***a informação*** a été extrait de l'article n° 1, paragraphe n° 4

M. Le Guern considère que l'identification des syntagmes nominaux devrait être réalisée par le numéro d'article et par le numéro de la ligne, soit où il est, soit où il commence. Certes, cette procédure est plus précise. Or, pour cette expérimentation, la numérotation adoptée, bien que moins précise, n'a guère compromis les résultats car les articles n'étant pas longs, des paragraphes entiers étaient presque toujours visibles sur l'écran. Pour une application professionnelle cependant, l'adoption d'une technique précise de manière à identifier les syntagmes nominaux est désirable, soit au moyen de la numérotation des lignes, soit d'une autre façon quelconque qui puisse les distinguer lorsque les articles sont présentés à l'écran.

3. Extraction des syntagmes nominaux

Pour l'extraction des syntagmes nominaux d'un corpus il a fallu d'abord les identifier. Pour cela, on a utilisé une démarche logico-sémantique, étant donné qu'il n'y a pas d'analyseur morpho-syntaxique ni des règles pour cette identification.

Le Guern montre (*Français Moderne*, juin 1991) qu'un syntagme nominal est le résultat de la mise en oeuvre de deux organisations logiques différentes. Si on prend comme exemple le syntagme nominal « o sistema de informação » afin de l'adapter à l'explication de Le Guern, on verra que : le mot « sistema », en tant qu'élément du lexique de la langue, ne désigne aucun objet quel qu'il soit, mais uniquement un ensemble de propriétés, sans la prise en compte d'un univers donné. Il relève d'une logique intensionelle, logique sans

référentiel et sans classe, constitué de relations et de propriétés envisagées indépendamment de quelque objet que ce soit. Dans ce cas, le mot « sistema » trouvera différentes acceptions dans plusieurs domaines. Par contre le terme « sistema de informação » prend sa valeur dans un univers précis du discours. C'est un prédicat lié. Le terme fait une référence au domaine de sciences de l'information. C'est le basculement de la logique intensionnelle à la logique extensionnelle. C'est la mise en relation des mots et des choses. Lorsque un déterminant opère sur ce terme, « o sistema de informação », on a le syntagme nominal. Dans ce raisonnement, on identifie aussi le centre du syntagme, c'est-à-dire « sistema », qu'est un prédicat libre.

3.1 Mise en forme des syntagmes nominaux

Afin d'éviter les problèmes de tri et en conséquence l'imprécision des résultats pendant la recherche d'information, on a adopté les dispositions suivantes pour la mise en forme des syntagmes nominaux :

a) saisie des syntagmes nominaux en caractères minuscules

Les mots apparaissent dans les textes sous plusieurs formes, soit commençant avec une lettre majuscule, soit tout en minuscule ou encore tout en majuscules. Les systèmes de gestion de bases de données utilisent le code de chaque lettre pour trier les mots. En conséquence, les résultats sont influencés puisqu'on ne peut pas éventuellement retrouver un mot qu'existe dans la base étant donné qu'il est écrit ou enregistré de différentes façons. Ainsi, il a fallu décider pour la conversion de tous les caractères des syntagmes nominaux en minuscules.

b) suppression des accents et des cédilles

Cette mesure a été adoptée car il y avait quelques fautes d'orthographe dans les articles choisis. Ainsi on a trouvé les mêmes syntagmes nominaux sous deux formes dont la seule différence était l'accent ou la cédille. D'autre part, ces signes sont presque souvent une source d'erreurs lorsque un utilisateur saisi une

recherche. En plus, les claviers n'utilisent guère une même norme de disposition des touches. Il y a des claviers du type américain, français et d'autres.

3.2 Calculs des syntagmes nominaux

Plusieurs situations se présentent dans les textes où le repérage des syntagmes nominaux n'est pas toujours évident. Cela arrive soit parce qu'il y a des éléments anaphoriques, soit parce qu'il y a des ellipses, ou soit parce qu'il y a d'autres situations où les syntagmes nominaux se trouvent cachés ; il est possible d'autre part, de trouver des syntagmes nominaux qui ne portent pas d'information. Ainsi, il a fallu adopter quelques règles afin d'extraire les syntagmes nominaux de façon homogène :

1. Syntagmes nominaux vides

   Par principe les articles sont composés de sections et de parties dont les titres ont été considérés comme étant des syntagmes nominaux. Or, on s'est vite rendu compte que plusieurs de ces syntagmes ne portaient pas d'information, comme par exemple : Conclusão (Conclusion), Objetivo (Objectif), Antecedentes (Antécédents), Introdução (Introduction), etc. On les a alors supprimés de la liste des syntagmes nominaux.

   On a trouvé également dans les textes des syntagmes vides, tels que : nesse sentido (dans ce sens), nesse contexto (dans ce contexte), uma vez (une fois que...), tal processo (un tel processus...), outro angulo (sous un autre angle), o momento (à ce moment...), etc. Ces syntagmes ont été aussi supprimés.

2. Syntagmes nominaux cachés dans des phrases avec factorisation

   L'extraction des syntagmes nominaux dans des phrases avec factorisation n'est pas toujours évidente, sauf quand on a une indication claire du syntagme comme par exemple dans la phrase suivante :

   o processo de negociação dos setores privado e público

Dans ce cas, le syntagme nominal de niveau un est clairement distingué comme étant ***os setores privado e público***, parce que le mot ***setores*** est au pluriel et il fait référence aux deux mots - ***privado e publico,*** au singulier - simultanément.

Par contre, on a trouvé des situations où on a eu du mal à identifier le syntagme nominal de manière précise. Dans ces cas, on a décidé d'extraire le syntagme nominal composé par chaque mot de la suite coordonné et le complément de la phrase. Exemples :

a) Le syntagme nominal : ***a análise, interpretação, avaliação e comunicação da informação pelos meios convenientes***

a donné les syntagmes nominaux suivants :

⇒ ***a análise da informação pelos meios convenientes***

⇒ ***a interpretação da informação pelos meios convenientes***

⇒ ***a avaliação da informação pelos meios convenientes***

⇒ ***a comunicação da informação pelos meios convenientes***

b) Le syntagme nominal : ***o potencial de conhecimento e inteligência da organização***

a produit les syntagmes nominaux suivants :

⇒ ***o potencial de conhecimento da organização***

⇒ ***o potencial de inteligência da organização***

Il existe également une autre forme de factorisation où les mots coordonnés apparaissent entre parenthèses comme un complément discriminatoire du terme ou de la phrase qui précède la parenthèse. Exemple :

La construction : ***rapidez e profundas transformações (políticas, econômicas, sociais, tecnológicas)***

a produit les syntagmes nominaux :

⇒ ***rapidez e profundas transformações políticas***

⇒ ***rapidez e profundas transformações econômicas***

⇒ ***rapidez e profundas transformações sociais***

⇒ ***rapidez e profundas transformações tecnológicas***

Cependant, il faut faire attention à la construction dans la parenthèse, parce qu'on ne peut adopter cette règle que pour les suites de mots coordonnés. Certes, il y a des constructions entre les parenthèses qui sont des phrases explicatives. Dans ces cas, les syntagmes nominaux qui apparaissent dans les parenthèses vont être extraits indépendamment. Or, malgré cette solution apparemment facile lors d'une extraction manuelle, elle n'apparaît pas réalisable dans une procédure automatique.

3. Phrases entre guillemets

Les guillemets sont utilisés normalement en deux situations : soit pour distinguer une citation soit pour distinguer un mot ou un terme (groupe limité de mots). Dans la situation d'une citation, on a fait l'extraction des syntagmes nominaux comme dans un texte normal, tandis que pour la deuxième situation on a simplement enlevé les guillemets. On a trouvé, cependant, des situations où le terme dans les guillemets a été identifié comme un syntagme nominal. Exemple : *a denominação de « Economia da Informação »*

Une fois, encore, on peut trouver des difficultés dans une procédure automatique d'extraction de syntagmes nominaux.

4. Phrase entre tirets

Les phrases entre tirets ont été traités de la même façon que les phrases entre parenthèses. La situation est similaire.

5. Article Zéro

Différemment de la langue française, en portugais on trouve souvent des phrases où les articles sont omis, donc des phrases qui n'ont pas des déterminants. À cet égard Cintra et Cunha observent :

> *« A la rigueur, il ne s'agit pas proprement, dans ce cas et dans les suivants d'omission d'article indéfini, mais de cas où il n'a jamais été employé de manière régulière.*

> *« Dans la phase primitive des langues romanes, l'article indéfini était d'utilisation restreinte. Avec le temps, ce déterminatif s'est introduit dans plusieurs constructions et, aujourd'hui, les diverses nuances de son emploi constituent une inestimable richesse stylistique pour chacune d'entre elles.*

> *« Par contre, nos grammairiens refusent cette généralisation et valorisation progressive de l'article indéfini, qu'ils ne voient qu'une simple et superflue influence du français, et où en fait peu sont actuellement les interdictions d'utilisation de ce déterminatif. Mais une telle discussion s'est révélée inutile parce qu'il ne s'agit pas d'un simple gallicisme possible d'être extirpé, mais d'une tendance générale des idiomes latins en recherche de formes plus expressives, avec plus de clarté et fermeté pour l'énoncé. »*[26]

Selon Le Guern, l'omission des articles indéfinis est plus courante dans les cas des substantifs abstraits au pluriel, comme par exemple : informações científicas, sistemas, etc.

---

[26] CUNHA, Celso et CINTRA, Lindsey. Nova Gramática do Português Contemporâneo. Lisboa : Edições João Sá da Costa, 1991. p. 242

« Em rigor, não se trata propriamente, nesses casos e nos seguintes de omissão do artigo indefinido, mas de casos onde ele nunca se empregou de forma regular.

« Na fase primitiva das línguas românicas, o artigo indefinido era de uso restrito. Com o correr do tempo, esse determinativo foi-se introduzindo em numerosas construções e, hoje, os variados matizes do seu emprego constituem uma inestimável riqueza estilistica de todas elas.

« Contra essa generalização e valorização progressiva do indefinido se manifestaram sempre os nossos gramáticos, que nela vêem uma simples e desnecessária influência do francês, onde, em verdade, poucas são actualmente as interdições ao uso do determinativo em causa. Mas tal guerra tem-se revelado inútil, e inútil precisamente porque não se trata, no caso, de um mero galicismo extirpável, e sim de uma tendência geral dos idiomas neolatinos em busca de formas mais expressivas, de maior clareza e vigor para o enunciado. »

Cintra et Cunha présentent également quelques situations où les articles sont omis :

- en cas d'énumération
  - par accumulation

Exemple : *perspectivas do agente da informação no contexto brasileiro: problemas, barreiras e desafios*

dont on trouve les syntagmes nominaux suivants :

⇒*perspectivas do agente da informação no contexto brasileiro*

⇒*o agente da informação no contexo brasileiro*

⇒*a informação*

⇒*o contexto brasileiro*

⇒*problemas*

⇒*barreiras*

⇒*desafios*

- par dispersion

Exemple : *políticas, procedimentos, diretrizes e sistemáticas para a organização da função*

dont les syntagmes nominaux sont :

⇒*políticas para a organização da função*

⇒*procedimentos para a organização da função*

⇒*diretrizes para a organização da função*

⇒*sistemáticas para a organização da função*

- on peut supprimer l'article défini lorsque le substantif est abstrait ou dans un proverbe, ou dans des phrases de comparaison brève.

  Exemple : *Conhecimento como recurso estratégico empresarial*

  dont les syntagmes nominaux sont :

  ⇒ *conhecimento como recurso estratégico empresarial*

  ⇒ *recurso estratégico empresarial*

- avant les mots qui indiquent des disciplines d'étude utilisées avec les verbes aprender, estudar, cursar, ensinar, et synonymes.

  Exemple : Aprender **Francês**, Cursar **Direito**

  dont les syntagmes nominaux sont :

  ⇒ *francês*

  ⇒ *direito*

6. Calculs des anaphores

Les éléments anaphoriques, en portugais, apparaissent souvent au moyen des particules suivantes : pronoms possessifs, pronoms démonstratifs, pronoms personnels, etc.

L'extraction des syntagmes nominaux cachés par les éléments anaphoriques n'a toujours pas été facile. Lorsque les sources de ces éléments étaient près d'eux, on a pu les résoudre facilement. Par contre, quand leurs sources se présentaient dans les paragraphes précédents, ou encore plus loin, l'extraction des syntagmes nominaux devenait très difficile. Malgré les difficultés rencontrées, on a quand même essayé de les résoudre.

Deux cas d'anaphores cependant n'ont pas pu être résolus : d'une part les anaphores sans sources, tels que : *nesse sentido* (dans quel sens? il n'y a pas de source dans le texte), *desse modo* (de quelle façon? il n'y a pas de source dans le texte), *nossa experiência* (quelle expérience? celle de l'auteur? celle

de techniciens d'information?), *etc.* Et pourtant, il a été facile de constater que ces syntagmes ne portent aucune information, et qu'ils sont plutôt des termes accessoires dans le processus d'écriture.

Le deuxième cas d'anaphore non résolu est celui des anaphores sans sources explicites, mais qui portent des informations du genre : *esse período pré-industrial, esse sistema de comunicação, aqueles benefícios que não podem ser mensurados monetáriamente, etc.* Dans ces cas, les syntagmes ont été conservés et transcrits tels quels, sans un quelconque traitement.

Bien qu'on ait résolu dans la plupart des cas les problèmes d'anaphores, les phrases résultantes sont parfois curieuses, comme par exemple :

⇒ *uma categoria de clientes conscientizados dos seus direitos a produtos e serviços de alta qualidade*

dont la solution est :

⇒ *uma categoria de clientes conscientizados dos direitos dos clientes conscientizados a produtos e serviços de alta qualidade*

Une manière de résoudre ce problème serait de remplacer les éléments anaphoriques seulement au moment de l'extraction des syntagmes qui les enveloppent. Ainsi, l'exemple ci-dessus reste : (où SN = syntagme nominal et le numéro qui représente le niveau du syntagme)

SN 4 : *uma categoria de clientes conscientizados dos seus direitos a produtos e serviços de alta qualidade*

SN 3 : *os direitos dos clientes conscientizados a produtos e serviços de alta qualidade*

SN 2 : *os clientes conscientizados a produtos e serviços de alta qualidade*

SN 1 : *produtos e serviços de alta qualidade*

Cette solution n'a pas été adoptée dans ce travail puisqu'on a préféré garder les syntagmes nominaux entièrement développés.

7. Calculs des ellipses

Le problème lié à ce type de figure est toujours dépendant de la capacité de se rendre compte de ce qu'il manque un quelconque mot dans une phrase. Il faut toujours analyser non seulement les phrases précédentes, mais aussi les phrases suivantes. Exemple :

⇒ *uma visão de longo prazo que assegure não só a sobrevivência (?), como também o crescimento da organização*

Quel est le complément du terme *sobrevivência*, c'est-à-dire, la survie de qui? La solution se trouve dans la phrase suivante : *o crescimento da organização*. Ainsi, le syntagme complet est :

⇒ *uma visão de longo prazo que assegure não só a sobrevivência da organização, como também o crescimento da organização*

Dans une procédure manuelle il n'y a pas de problèmes pour trouver la solution des ellipses; par contre, dans une procédure automatique on trouvera sûrement des difficultés pour résoudre ce type de figure.

3.3 Typologie des syntagmes nominaux trouvés

Étant donné l'inexistence d'analyseur pour l'extraction automatique de syntagmes nominaux d'un corpus en langue portugaise, on l'a fait manuellement et en utilisant une

démarche logico-sémantique. Du fait que les langues française et portugaise sont de même origine et possèdent une structure grammaticale analogue, on a utilisé comme base les travaux déjà développés pour la langue française par le groupe SYDO[27].

Les syntagmes nominaux extraits se présentent sous diverses combinaisons d'éléments syntaxiques. Pour connaître de quelle façon ces éléments ont été rassemblés, une typologie de ces syntagmes s'impose; pour cela, on utilisera la notation BNF - Backus Norm Form ou Backus-Naur Form. Exemple :

<prédéterminante article> ::= a | as | o | os | um | uma |

où :

' ::= ' - signifie « est définit comme »

' <prédéterminante article> ' - éléments syntaxiques ou non-terminaux

' | ' - ou

ensemble de symboles sans ' < ' et ' > ' - éléments terminaux

Les typologies des syntagmes nominaux ont été faites en groupant les syntagmes nominaux trouvés selon la ressemblance de leurs éléments syntaxiques. Ainsi, la description faite ici, ne reflète que les syntagmes nominaux trouvés. Ces typologies sont plutôt un ensemble de descriptions qui peuvent un jour servir à former un ensemble de règles. Le développement de cet ensemble de règles n'est pas l'objet du présent travail, mais fait partie de mon programme personnel pour ma thèse de doctorat.

1. <syntagme nominal simple> ::= <prédéterminante article> <centre de syntagme> | <prédéterminante article> <centre de syntagme adjectivé>

[27] Dont on distingue le travail de Jean-Paul Metzger, *Syntagmes Nominaux et information textuelle : reconnaissance automatique et représentation.* Lyon, 1988. 324 p. Thèse de doctorat d'Etat, Université Claude Bernard, Lyon I.

<déterminante> ::= <prédéterminante article> | <prédéterminante indéfini> | <prédéterminante quantitatif>

<prédéterminante article> ::= a | o | as | os | um | uma

<prédéterminante indéfini> ::= muitos | muitas | alguns | algumas ...

<prédéterminante quantitatif> ::= <numéral> | <nombre> |

<prédeterminante article> <numéral> | <nombre> <numéral cardinal multiple de mille>

<numéral> ::= <numéral cardinal> | <numéral ordinal>

<numéral cardinal> ::= dois | três | quatro | ...

<numéral cardinal multiple de mille> ::= mil | milhão | bilhão | ...

<numéral ordinal> ::= primeiro | segundo | terceiro |...

<nombre> ::= <num>

<num> ::= <num> <chiffre>| <chiffre>

<chiffre> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

<centre de syntagme> ::= <substantif>

<substantif> ::= <nom commun> | <nom propre> |

<nom commun> ::= <nom commun concret> | <nom commun abstrait>

<nom commun concret> ::= biblioteca | livro | computador ...

<nom commun abstrait> ::= informação | riqueza | estratégia ...

<nom propre> ::= França | Brasil | João | Pedro | Maria | Paris ...

<centre de syntagme adjectivé> ::= <adjectif> <nom commun> | <nom commun> <adjectif> | <nom commun> <adjectif 1> <adjectif 2>

<adjectif> ::= alta | estratégica | empresarial | industrial | europeu ...

Remarque.: les notations <adjectif 1> et <adjectif 2> sont numérotées pour indiquer qu'ils signalent deux adjectifs distincts.

Exemples :

- *a informação, as informações, os produtos, os serviços, os países, a ação, o Brasil, os Estados Unidos, a França,* etc.
- *a administração estratégica, a atividade empresarial, a atividade industrial, a abordagem estratégica, a alta direção, a alta administração, os engenheiros japoneses, o Mercado Comum Europeu, o plano estratégico, o caso brasileiro,* etc.

2. <syntagme nominal simple 1> ::= <prédéterminante indéfini> <centre de syntagme> | <prédéterminante indéfini> <centre de syntagme adjectivé>

   Exemple : *muitas universidades, muitos autores, muitos usuários, alguns autores, algumas universidades,* etc.

3. <syntagme nominal simple 2>[28] ::= <prédeterminante quantitatif> <centre de syntagme>

   Exemple : *o primeiro mundo, 15000 empresas, 100 mil titulos, etc.*

4. <syntagme nominal type date> ::= <date>[29]

   <date> ::= <année> | <mois> de <année>

   <année> ::= 19<chiffre><chiffre>

---

[28] Cette description n'est pas exaustive car on n'a décrit que les cas trouvés dans le corpus

[29] Idem

<chiffre> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

<mois> ::= janeiro | fevereiro | março | abril | maio | junho | julho | agosto | setembro | outubro | novembro | dezembro

Exemples : *1914, 1947, 1987, 1991, novembro de 1990,* etc.

5. <syntagme nominal composé> ::= <déterminante> <centre de syntagme> <syntagme prépositionnel> | <déterminante> <centre de syntagme adjectivé> <syntagme prépositionnel> |

   <syntagme prépositionnel> ::= <syntagme prépositionnel> [<préposition> <syntagme nominal simple 1> | <contraction prépositionnelle> <centre de syntagme> | <contraction prépositionnelle> <centre de syntagme adjectivé>] | <préposition> <syntagme nominal simple 1> | <contraction prépositionnelle> <centre de syntagme> | <contraction prépositionnelle> <centre de syntagme adjectivé>

   <préposition> ::= de | para | com | em | por | ...

   <contraction prépositionnelle> ::= do | da | dos | das | no | nas | pelo | pela | pelos | pelas

   Exemples : *a análise da informação, a análise da eficiência, a ação do sistema, a aceitação generalizada do conceito, o objetivo do sistema, a equipe da organização,* etc.

6. <syntagme nominal avec article zéro> ::= <déterminante> <centre de syntagme> <préposition> <nom commun abstrait pluriel>

   <nom commun abstrait pluriel> ::= riquezas | estratégias | informações ...

   Remarque.: On a suivi ici la discussion présentée dans la section précédente (3.3, cas n°. 5 - Article Zéro)

Exemples : *a acumulação de riquezas, a adoção de estratégias*, etc.

7. <syntagme nominal avec expansion prépositionnelle> ::= <déterminante> <centre de syntagme> <préposition> <nom commun> | <déterminante> <centre de syntagme> <préposition> <nom commun> <adjectif> | <déterminante> <centre de syntagme> <préposition> <nom commun> <préposition> <nom commun>

   Exemples : *a análise de conteúdo, o sistema de informação, a característica de valor de troca, a gestão de portfólio empresarial*, etc.

8. <syntagme nominal avec proposition relative déterminative> ::= <déterminante> <centre de syntagme> <proposition relative déterminative> [30]

   Exemples :

   ⇒ *a empresa que não tiver meios materiais ;*

   ⇒ *os potenciais de conhecimento que devem ser identificados dentro da estrutura de serviços tecnológicos ;*

   ⇒ *a ciência interdisciplinar que investiga as propriedades da informação ;*

   ⇒ *os sistemas que participam do desenvolvimento da inteligência competitiva da empresa ;*

   ⇒ *uma expressão que designa sistematizações relacionadas com as evoluções e mutações que marcaram a economia dos países desenvolvidos ;*

   ⇒ *algumas formulas que conduzem ao crescimento econômico e a inovação*

---

[30] Cette typologie ne s'applique pas aux propositions rélatives déterminatives explicatives.

9. <syntagme nominal appellatif> ::= <prédéterminante article> <centre de syntagme> <nom propre> | <prédéterminante article> <centre de syntagme> <sigle> | <prédéterminante article> <centre de syntagme> <préposition> <nom propre> | <prédéterminante article> <centre de syntagme> <sigle> <préposition> <nom propre>

   Exemples : *o chinês Sun Tzu, o ciclo PDCA de Shewart*, etc.

10. <syntagme nominal avec double rection> ::= <déterminante> <centre de syntagme> <contraction prépositionnelle> <centre de syntagme> <contraction prépositionnelle> <centre de syntagme> | <déterminante> <centre de syntagme> <contraction prépositionnelle> <centre de syntagme adjectivé> | <déterminante> <centre de syntagme adjectivé> <contraction prépositionnelle> <centre de syntagme adjectivé> | <déterminante> <centre de syntagme adjectivé> <contraction prépositionnelle> <centre de syntagme>

    Exemples :

    ⇒ *a adaptação da empresa às mudanças ambientais,*

    ⇒ *a avaliação da informação pelos meios convenientes,*

    ⇒ *o armazenamento da informação nas organizações*, etc.

11. <syntagme nominal nom propre> ::= <nom propre>

    Exemple : *Toffler, Demin, Toyoda*, etc.

    Remarque. : Du fait que les syntagmes nominaux constitués de noms propres et les syntagmes nominaux constitués de dates semblent posséder des caractéristiques propres, on les distingue de tous les autres.

4. Conclusion

Pendant la procédure d'extraction des syntagmes nominaux du corpus, bien qu'on ait pris toutes les précautions, ça a été difficile de garder une certaine homogénéité. Les raisons principales sont : (a) la diversité de style de rédaction des articles, étant donné qu'ils ont été écrits par quinze auteurs différents. On s'est rendu compte qu'il y a un lien étroit entre la facilité d'extraction des syntagmes nominaux et la clarté des articles ; (b) un ensemble de règles pour orienter plus précisément la procédure d'extraction des syntagmes nominaux a fait défaut ; (c) le processus d'extraction n'a pas été continu, car des événements tels que les examens (écrit et oral) à l'Ecole et l'élaboration de la note de synthèse ont interrompu le travail. Ainsi, il a fallu faire une révision des syntagmes nominaux extraits, au fur et à mesure du chargement de la base de données et de la construction de l'arborescence.

Les résultats de cette étape sont consolidés dans la figure 1, où on trouve le nombre de syntagmes nominaux avec et sans doublons, le nombre de mots, de paragraphes, de lignes et de pages dans chaque article.

Le nombre de syntagmes nominaux avec doublons tient compte de la multiplicité d'occurrence d'un même syntagme nominal dans un article donné, tandis que le nombre de syntagmes nominaux sans doublons ne le fait pas. On a constaté par ailleurs qu'il y a des doublons qui peuvent apparaître ou non sur l'ensemble des syntagmes nominaux appartenant à plus d'un article différent. La colonne des syntagmes nominaux sans doublons ne tient pas compte de cet aspect ; le nombre total des syntagmes nominaux par niveaux et sans doublons se retrouvent dans le chapitre III, section 5.

A partir de ce tableau et en tenant compte du nombre de syntagmes nominaux avec doublons (cette variable a été prise parce qu'elle représente la totalité de syntagmes nominaux d'un article), on arrive aux moyennes suivantes :

Nombre moyen de syntagmes nominaux par lignes = 2,88

Nombre moyen de syntagmes nominaux par paragraphe = 1,17

Nombre moyen de syntagmes nominaux par page[31] = 139,96

Nombre moyen de mots par syntagmes = 3,21

Article	Nombre de Syntagmes		Nombre de			
	avec doublons	sans doublons	mots	paragraphes	lignes	pages
1	681	500	1.904	38	199	4
2	576	484	1.769	49	196	4
3	555	362	1.834	80	245	5
4	502	344	2.010	55	236	5
5	726	590	1.867	77	234	5
6	588	403	1.960	36	196	4
7	624	475	2.106	28	191	4
8	642	484	2.004	45	218	4
9	590	471	1.838	30	179	4
10	479	353	1.560	32	154	3
11	467	334	1.709	33	167	3
12	611	502	2.027	80	257	5
13	519	406	1.460	42	168	4
14	672	499	2.380	42	235	5
15	586	428	1.909	33	178	4
Total	8.818	6.635	28.337	700	3.053	63

Figure 1. Tableau résumé du corpus d'articles servant à la construction de la base de données

On a utilisé la moyenne bien que l'on sache qu'il s'agit d'une mesure d'utilité douteuse pouvant être facilement influençable par une valeur trop grande (ou trop petite) le faisant perdre complètement sa représentativité. Le but de la présentation du tableau des moyennes n'est pas de chercher une relation entre les variables, mais de décrire le travail d'extraction des syntagmes nominaux.

Concernant les moyennes obtenues on considère que :

a) La taille des paragraphes varie largement. On trouve des paragraphes avec une seule ligne et d'autres avec dix lignes ou plus. Cet aspect dépend du style de rédaction des auteurs ;

[31] Les donnés descriptives des articles ont été pris dans des fichiers dont les caractères sont en Times New Roman, taille 11.

b) En ce qui concerne la taille des pages, on peut dire qu'elle présente presque le même problème, étant donné que les dernières pages de chaque article ne sont pas toujours pleines ;

c) A propos du nombre moyen de syntagmes nominaux par ligne, bien que l'on trouve des lignes remplies à moitié, elles apparaissent en générale complètes ;

d) Au sujet du nombre moyen de mots par syntagmes nominaux, on peut croire qu'il représente plutôt les syntagmes de premier et de deuxième niveau et à la rigueur ceux de troisième niveau, mais en aucun cas il s'agirait des syntagmes de quatrième et de cinquième niveau. Ces derniers sont très longs, ayant quelquefois vingt mots ou plus. Ainsi, pour obtenir une valeur plus fiable il aurait fallu faire ce calcul plutôt dans chaque niveau.

Pour une analyse plus approfondie et envisageant de trouver des relations entre les variables présentées dans la figure 1, un travail spécifique mérite d'être faite en considérant des critères plus orientés à ce type d'étude.

En ce qui concerne la procédure d'extraction des syntagmes nominaux, on a trouvé quelques points assez curieux qui se prêtent à des études plus approfondies. Ils sont :

a) comment traiter les articles zéro dans une procédure automatisée d'extraction des syntagmes nominaux ;

b) l'automatisation des calculs des anaphores ;

c) l'automatisation des calculs des ellipses ;

d) l'automatisation de l'extraction des syntagmes nominaux dans les phrases utilisant la factorisation dans leur construction. On a décidé de faire le calcul des images logico-sémantiques (terme utilisé par Omar Larouk, en sa thèse de

doctorat[32], pour exprimer la procédure de détermination de toutes les solutions liées à une connexion à l'intérieur d'une phrase). On a eu des difficultés pour trouver la bonne solution, excepté lorsqu'on connaissait le contexte où les éléments syntaxiques étaient évidents. Ainsi, dans les cas où on ne trouvait pas cette situation, on a décidé de prendre comme solution tous les syntagmes nominaux résultants du calcul.

Bien qu'on ait adopté des solutions pour chacun de ces points, l'analyse et formalisation de solutions définitives est indispensable. Cela sera fait dans le cadre de ma thèse de doctorat.

[32] Omar LAROUK. Extraction de connaissances à partir de documents textuels : traitement automatique de la coordination (connecteurs et ponctuation). Université Claude Bernard - Lyon I, Thèse de Doctorat, 1994.

« Il est impossible de ranger les pièces à qui n'a une forme du total en sa tête. A quoi faire la provision des couleurs à qui ne sait ce qu'il a à peindre. »

MONTAIGNE (1533 - 1592)

Chapitre II

Développement de la maquette

1. Considérations préliminaires

Tout d'abord, pour la construction d'une maquette d'un système de recherche d'information, il a fallu connaître quelles étaient les caractéristiques de l'interface de ce système et comment il devrait fonctionner. On a déjà consigné, sur la Note de Synthèse, les caractéristiques souhaitables à une telle interface :

> *« Améliorer la convivialité d'utilisation de ces systèmes [systèmes de recherche d'information] et augmenter la précision des résultats de la recherche d'information sont les traits principaux. Ainsi on peut envisager les caractéristiques souhaitables suivantes :*
>
> - *interface de dialogue en langage naturel*
>
> *Cette interface doit posséder une interaction facile et naturelle avec l'usager de manière à éviter les contraintes de l'apprentissage d'un langage avec des règles rigides de grammaire. Pour cela, il faut donc utiliser les éléments que les humains utilisent pour se communiquer les uns avec les autres. Autrement dit, il faut utiliser un langage plus proche possible du langage parlé par les humains. Ainsi l'interface doit être construite vers l'usager final ;*

- *méthode de traitement du contenu des documents et des requêtes en considérant la façon dont les humains reconnaissent la pertinence ou non d'un document*

C'est-à-dire, il ne suffit pas de faire la simple extraction de mots, groupes de mots ou structures syntaxiques pour représenter les contenus des documents. Mais, principalement identifier " les parties plus petites du discours (documents) que peuvent servir de base à une relation référentielle autonome, les syntagmes nominaux" [33].

Les humains sont capables de distinguer en un coup d'oeil lorsqu'ils regardent un document, s'il est ou non pertinent par rapport à son besoin d'information. Certes derrière ce processus il y a des connaissances qu'ont été accumulées pendant des années. Or, lorsque les hommes font le choix des documents pertinents parmi les autres, en regardant soit les titres, soit le propre contenu, ils ne font que comparer les éléments référentiels existant dans ces champs avec ceux qu'ils ont définis comme leur besoins d'information ;

- *mécanisme de retro-alimentation*

Pour améliorer encore la précision des résultats de la recherche d'information, on considère qu'un mécanisme d'aide aux usagers sera intéressant, pour les raisons suivants: a) permet un raffinement de la requête; b) ce type d'interaction entre les usagers et les systèmes, permettra une meilleure connaissance de la base de données et du domaine de ses documents. Il y a déjà quelques prototypes qui ont développé ce type de mécanisme, appelé reformulation [voir MECABOUCHE].

En outre, à mon avis l'utilisation de plusieurs modes d'expression peut rendre les SRI [Systèmes de Recherche de Information] plus conviviaux et agréables pour les usagers. »[34]

---

[33] Michel LE GUERN. « Un analyseur morpho-syntaxique pour l'indexation automatique », *Le Français Moderne*, Juin, 1991, t. LIX, n°. 1, p. 24.

[34] Hélio KURAMOTO. Les Systèmes de Recherche d'Information en Langage Naturel. Note de Synthèse, Avril, 1995. p. 36-37.

De cette manière et en accord avec la suggestion du M. Le Guern, on a travaillé sur le développement d'une interface où la procédure de recherche d'information soit le résultat d'une interaction étroite entre le système et l'utilisateur. M. Le Guern montre la façon dont un outil d'aide à la recherche d'information pourrait fonctionner :

> *« ... L'outil d'aide à l'interrogation donne, pour chaque syntagme d'un niveau donné, la liste des syntagmes de rang immédiatement supérieur qui le contiennent, avec le nombre d'occurrences de chaque syntagme dans le corpus. L'utilisateur peut ainsi cheminer dans l'ensemble des descripteurs, en réduisant progressivement le nombre d'occurrences, jusqu'au moment où il obtiendra la quantité de références correspondent à l'ordre de grandeur qu'il souhaite. »*[35]

L'idée principale cherche à faire le système naviguer dans l'arbre des syntagmes nominaux à partir d'une demande de l'utilisateur. A partir de cette demande le système cherche et présente tous les syntagmes nominaux de rang immédiatement supérieur à l'utilisateur, afin qu'il puisse faire son choix et demander ensuite au système le raffinement de la recherche. Cette procédure est répétée jusqu'au moment où l'utilisateur atteint son besoin d'information ou lorsque le niveau le plus élevé de l'arborescence des syntagmes nominaux est atteint. La figure 2 illustre cette procédure.

Il y a deux sortes d'interface que l'on peut développer : 1) soit en utilisant le langage naturel pour maintenir une interaction libre entre le système et l'utilisateur ; 2) soit au moyen d'une interaction mixte, où le système tout d'abord offre aux usagers la possibilité de recherche et ensuite les aide à la raffiner. Cette procédure qui passe par l'utilisation des menus, laisse aux usagers la possibilité de choisir les syntagmes pertinents, de demander au système la présentation des documents trouvés, ou alors de recommencer la recherche.

---

[35] Michel LE GUERN. « Un analyseur morpho-syntaxique pour l'indexation automatique », *Le Français Moderne*. Juin. 1991, t. LIX, n°. 1, p. 34

La deuxième option a semblé plus appropriée pour l'implémentation d'une maquette puisque plus simple de développer dans l'espace de temps disponible qu'on a eu. Ainsi, tel a été le type d'interface adoptée pour ce travail.

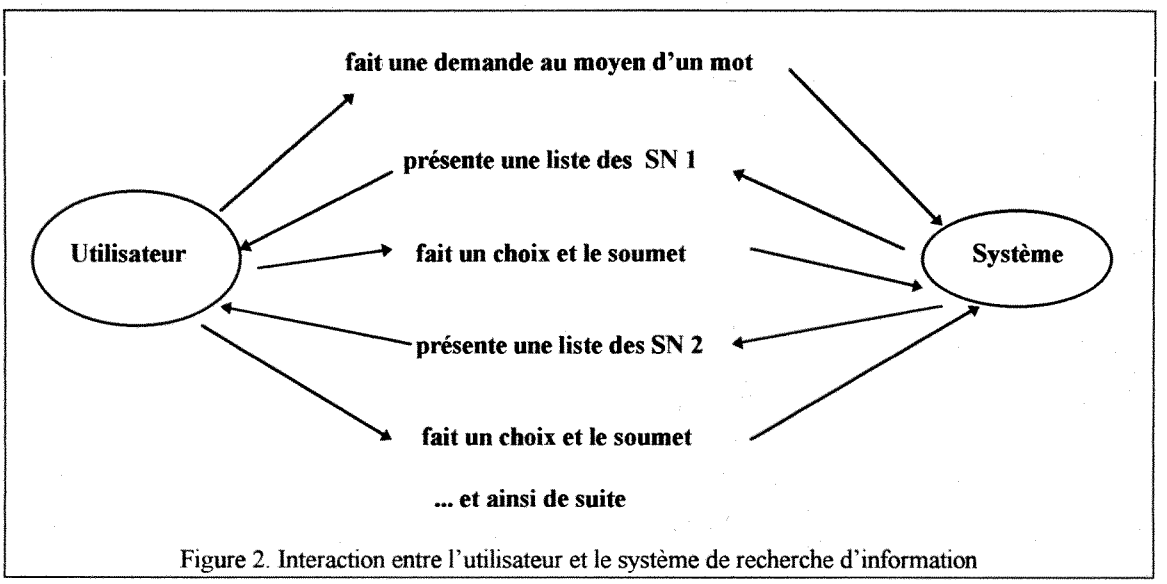


Figure 2. Interaction entre l'utilisateur et le système de recherche d'information

2. Choix de l'approche de développement de la maquette

Le développement d'une maquette peut être fait en utilisant plusieurs outils. Ces outils ont été choisis selon ce que l'environnement de la maquette imposait comme caractéristiques souhaitables. Les caractéristiques souhaitables à considérer sont alors : 1) souplesse dans la construction d'une maquette plus conviviale et de manipulation facile ; 2) flexibilité pour effectuer des modifications dans un court espace de temps ; 3) l'importation/exportation des données de/vers les fichiers en format Word 6.0a ; 4) la construction manuelle de l'arborescence ; 6) la disponibilité du logiciel.

Il y a fondamentalement deux grandes classes d'outils à analyser : 1) le développement de la maquette en utilisant la programmation conventionnelle, à partir de langages de programmation comme C++, Pascal etc. ; 2) le développement de la maquette utilisant un logiciel de gestion de bases de données.

L'approche de développement de la maquette en employant des langages de programmation de haut niveau (C++, Pascal, etc) possède l'avantage d'offrir les possibilités d'optimisation du temps de réponse et de construction de la maquette selon ce que l'on souhaite, rendant ainsi le système plus conviviale et sur mesure. Or cette approche demande un peu plus de temps en ce qui concerne la construction de la maquette, étant donné qu'il faut programmer toutes les routines, les procédures d'accès aux structures de données, les procédures de contrôle et de formatage d'écran, les procédures d'importation et d'exportation de fichiers, etc. Le temps qu'on a eu n'a pas été suffisant pour mettre en oeuvre une telle maquette. En plus, la mise en place des changements de manière rapide est pratiquement impossible dans les conditions existantes.

L'approche d'utilisation d'un système de gestion de bases de données, offre toutes les possibilités de création d'une maquette conviviale et de mise en place des ajustements nécessaires dans un court espace de temps. La compatibilité relative à l'échange de données (importation et exportation de données) avec le format Word 6.0a est en liaison avec le logiciel choisi. Il y a aujourd'hui plusieurs logiciels avec une telle caractéristique. D'autre part, les applications développées sur les systèmes de gestion de bases de données ne sont pas toujours performantes puisqu'elles sont trop génériques et demandent de gros ressources de l'ordinateur. C'est pourquoi Il est nécessaire de choisir le système de gestion de base de données plus approprié au type d'application à être développée.

Parmi les possibilités d'utilisation des systèmes de gestion de bases de données, on distingue principalement les systèmes de gestion de bases de données relationnelles et les systèmes de gestion de base de données textuelles. En principe les systèmes de gestion de bases de données textuelles (MINISIS, MicroIsis, Adhoc plus, Basis plus, BRS, etc.) sont plus adaptés au présent travail puisqu'ils possèdent des fonctions et des caractéristiques appropriés au traitement et à la recherche de l'information textuelle. Cependant, les interfaces de recherche et la procédure d'indexation sont déjà prêtes et n'admettent guère d'adaptation. C'est la raison pour laquelle on n'a pas pu travailler avec ces logiciels pour le développement de la maquette.

Ainsi, il ne restait que les systèmes de gestion de bases de données relationnelles, qui malgré les difficultés de performance et le fait qu'ils ne sont pas appropriés pour le traitement de textes, permettent le développement des applications selon les caractéristiques souhaitables.

Le logiciel Access (construit par Microsoft) a été choisi en considérant les critères déjà discutés en plus des facteurs suivants :

a) **Convivialité et souplesse de développement d'application**

- à cause de son interactivité et de son interface graphique, intégré à Windows, ce logiciel possède des outils d'aide et d'assistance à ceux qui développent les applications.

- les applications sont développées au moyen de la définition des paramètres, des macros, des procédures et d'un langage capable de gérer des procédures du type événementiel, etc.

b) **Compatibilité avec les logiciels Word 6.0a et Excel**.

- le logiciel Access possède une grande facilité pour importer et exporter des données, surtout avec les logiciels Word 6.0a et Excel entre autres formats.

d) **Rapidité de développement et de mise à jour d'application**.

- facteur le plus important et déterminant pour le choix de ce logiciel, étant donné la limitation du temps pour ce travail. Cette facilité permet de changer au fur et à mesure des besoin les applications de manière assez rapide.

Tous ces facteurs ont été constatés pendant l'utilisation de ce logiciel. Il faut souligné en plus qu'il s'agit d'un logiciel d'apprentissage facile. La principale contrainte par

contre, est la taille maximale d'un champ de 256 caractères. En conséquence, d'autres contraintes se produisent, parmi lesquelles l'impossibilité d'opérer des liaisons entre des champs ayant une telle taille.

3. Modèle de données relationnel

Le logiciel Access utilise l'approche des structures de données de type relationnel pour gérer les bases de données. Cette approche regroupe un ensemble de concepts tous déjà bien connus, mais que méritent d'être présentés une fois de plus. Les définitions ont été empruntés de l'ouvrage de M. M. Vetter, *Modélisation des données : approches globales et orientée objets*[36].

a) Tuple

Tuple c'est une liste de valeurs; une même valeur peut apparaître plusieurs fois.

Exemple : <5, a informação, 7, 8>

L'exemple ci-dessus signifie que : le syntagme nominal *a informação*, dont le code est 5, a été extrait de l'article 7, paragraphe 8.

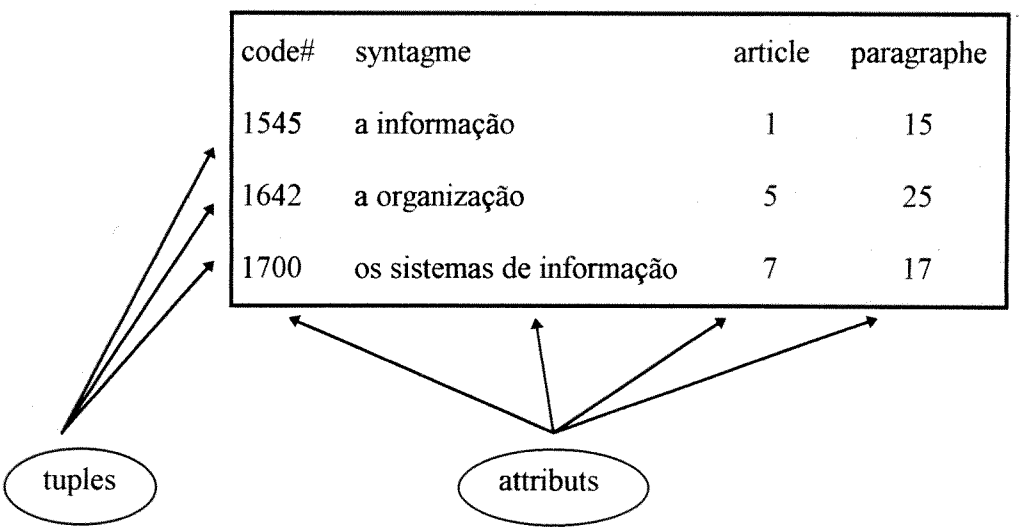
b) Relation

Relation c'est un ensemble de tuples. Une relation est normalement représentée par une table où les colonnes représentent les attributs - chaque colonne contient les valeurs d'un unique et seul domaine - et les lignes forment les tuples. Un certain tuple ne peut apparaître qu'une seule fois dans la relation.

---

[36] Max VETTER. Modélisation des données : Approches globale et orientée objets. Paris : Dunod Informatique, 1992

Exemple :



Figue 3. Exemple d'une relation

Les caractéristiques d'une relation :

- elle a un nom ;
- elle a de 0 à « n » tuples (lignes) ;
- elle a de 1 (une) à « m » colonnes aussi appelées attributs ;
- à l'intérieur d'une relation tous les attributs ont un nom unique ;
- les valeurs d'un attribut sont issues du même domaine, c'est-à-dire les valeurs d'un attribut ont les mêmes caractéristiques (numérique, alphabétique ou alphanumérique) ;
- une relation a au moins une clé ; la clé est un attribut ou un ensemble minimal d'attributs capables d'identifier de façon univoque les tuples de la relation ;
- chaque relation a une clé primaire ;

- parfois de situations se présentent où il est nécessaire l'utilisation de clés composées de plus d'un attribut ; c'est le cas, par exemple, des syntagmes nominaux avec double rection, où un même syntagme nominal est lié avec deux syntagmes de niveau 1 distincts. Ainsi, la clé doit être formé par les syntagme de niveau 2 et le syntagme de niveau 1, ce qui permet d'accéder au syntagme de deuxième niveau avec double rection, soit par une syntagme de niveau 1, soit par l'autre ;

Syntagme de Niveau 2	**Syntagme de Niveau 1**
a avaliação da informação pela empresa	a informação
a avaliação da informação pela empresa	a empresa
l'évaluation de l'information par l'entreprise	l'information
l'évaluation de l'information par l'entreprise	l'entreprise

Figure 4. Relation SN 2 - SN 1

Dans la relation de la figure 4, on voit le besoin de créer une clé composée par deux attributs puisqu'un même syntagme de niveau 2 peut être lié à plusieurs syntagmes de niveau 1.

- une relation peut être représentée de la manière suivante :

  R(code#, syntagme, article, paragraphe)

c) Normalisation des relations

Les relations non-normalisées présentent en général des données redondantes. Cette redondance provoque sur le système une plus grande occupation de la mémoire, en plus du risque d'anomalies de mémorisation. Ces anomalies sont des difficultés liées aux opérations d'insertion, de modification et de suppression, qui peuvent amener les relations vers un état ne correspondant pas à une description de la réalité.

Pour corriger ces problèmes, E.F. Codd[37,38] a été le premier à appliquer les règles de normalisation, en nombre de cinq aujourd'hui. Lorsqu'une relation respecte toutes les cinq règles, on l'appelle une relation entièrement normalisée. Pourtant, dans la pratique, on se contente des relations qui respectent les trois premières règles, celles qu'on appelle les relations en troisième forme normale (ces règles sont connues aussi comme 1FN - Première Forme Normale, 2FN - Deuxième Forme Normale, 3FN - Troisième Forme Normale, 4FN - Quatrième Forme Normale, 5FN - Cinquième Forme Normale).

a) Première Forme Normale

Dans une relation en première forme normale, les attributs ne peuvent prendre que des valeurs simples. Ainsi, dans l'intersection d'une colonne (attribut) et d'une ligne (tuple) il ne peut y avoir qu'une valeur.

Une autre forme de définir cette règle est : une relation est en première forme normale si tous les attributs non-clés sont dépendants fonctionnels de la clé. Un attribut est dit dépendant fonctionnel d'une clé si à chaque valeur d'une clé ne corresponde qu'une seul valeur de l'attribut.

b) Deuxième Forme Normale

Une relation en deuxième forme normale est caractérisée par le fait que tous ses attributs non-clés dépendent fonctionnellement de toute la clé (critère de 1FN) et non seulement d'une partie de la clé. C'est-à-dire, une relation est en 2FN si elle respecte la 1FN et que chaque attribut non-clé dépend pleinement fonctionnellement de la clé.

Ce critère s'applique aux relations dont la clé est composée d'un minimun deux attributs ayant au moins un attribut non-clé.

---

[37] E. F. CODD. « A relational model for large shared data banks ». *CACM*. 1970, vol. 13, n° 6.

[38] E. F. CODD. « Further normalization of the relational model ». *Data Base Systems, Courant computer science symposium 6*, 1971. Rustin R. Editeur, Englewood Cliffs, New Jersey 1972.

c) Troisième Forme Normale

Une relation est en troisième forme normale si elle respecte la 2FN et qu'elle ne porte aucune dépendance transitive (c'est-à-dire, il ne doit pas avoir de dépendance fonctionnelle entre des attributs qui ne sont pas des clés candidates). Une relation peut avoir d'autres clés en plus de la clé primaire, celles-ci étant des clés candidates, lesquelles à leurs tour sont en liaison simple les unes avec les autres.

Voici les principaux concepts pour la modélisation relationnelle de données. Il ne semble pas nécessaire de discuter plus en détail le modèle relationnel dans ce travail, car ce qui compte c'est la bonne utilisation de ce modèle. On présentera ainsi, par la suite, le modèle de données des syntagmes nominaux pour la construction de la maquette.

4. Modèle de données pour les syntagmes nominaux

Pour la modélisation de données, il est nécessaire tout d'abord de connaître le contexte du corpus et les faits qui caractérisent les syntagmes nominaux. Il faut également tenir compte des considérations présentées dans la section 1 de ce chapitre du fait qu'elles correspondent non seulement à l'approche de l'interface de recherche d'information, mais aussi à la démarche interactive entre l'utilisateur et la maquette de recherche d'information.

A la lumière du travail de construction de la base de données, on distingue l'existence de deux entités. Selon Vetter :

> *« Une entité est un exemplaire différenciable et identifiable d'une chose, d'une personne ou d'un concept concret ou abstrait, pour lequel on doit gérer des informations significatives. ...Il y a des auteurs qui considèrent qu'une association (par exemple, l'union d'une femme et d'un homme est aussi une entité). »*[39]

[39] Max VETTER. Modélisation des données : Approches globale et orientée objets. Paris : Dunod Informatique, 1992.

Ainsi, les deux entités sont :

a) ARTICLE (document du corpus) ;

b) SYNTAGME NOMINAL .

Les articles et les syntagmes nominaux constituent ensemble un contexte bien définit; pour la modélisation des données on a considéré les faits suivants :

1. Les articles sont numérotés en ordre séquentiel à partir de 1 ;
2. Pour chaque article les paragraphes sont aussi énumères en ordre séquentiel à partir de 1 ;
3. A chaque article correspond un titre d'une taille d'environ 256 caractères ;
4. A chaque article correspond un texte d'une longueur plus grande que 256 caractères ;
5. Un même syntagme nominal peut apparaître en plusieurs articles ;
6. Un même syntagme nominal peut apparaître en plusieurs paragraphes à l'intérieur d'un article donné ;
7. Les syntagmes nominaux peuvent être classés en cinq niveaux, selon le contexte de ce travail ;
8. Un même syntagme nominal peut être classé en plus d'un niveau ;
9. Il existe une association entre les syntagmes nominaux d'un niveau donné avec d'autres d'un niveau immédiatement supérieur (construction de l'arborescence) ;

10. Un syntagme nominal peut être associé à plusieurs syntagmes nominaux de niveau immédiatement supérieur ;

11. Plusieurs syntagmes nominaux peuvent être associés à un même syntagme nominal (le cas de double rection) ; ce syntagme à son tour peut appartenir à des niveaux distincts, ce qui dépend du niveau du syntagme nominal immédiatement inférieur ;

12. Il y a un centre de syntagme nominal associé à chaque syntagme nominal de premier niveau ;

13. Les syntagmes nominaux de premier niveau ont comme association de niveau inférieur les centres des syntagmes ;

14. Les mots associés aux syntagmes nominaux en dehors de l'ensemble des centres de syntagmes nominaux, sont aussi associés aux syntagmes de premier niveau en fonction de son importance dans la recherche d'information ;

15. Les centres des syntagmes nominaux possèdent des flexions en genre et en nombre.

On définira quelques termes que seront utilisés dans la construction du modèle de données de façon à éviter des confusions. Ainsi :

a) on utilisera le terme TABLE pour désigner un ensemble de tuples, au lieu de relation ;

b) et le terme RELATION pour désigner une association entre deux TABLES ou plus.

A partir des faits énumérés, on a conçu les structures de données nécessaires pour la construction de la maquette de recherche d'information. Toutes les tables conçues ont été

soumises aux règles de normalisation. Elles sont ainsi en 3FN. Ces tables seront explicitées selon la nomenclature suivante :

T (attr 1, attr 2, attr 3)

où : T - est le nom de la table et sera représenté en majuscules

attr 1 - l'attribut ou les attributs qui composent la clé ; il sera représenté en minuseules et sera souligné

attr 2, attr 3... - sont les noms des attributs que appartiennent à la table ; ils seront représentés en minuscules

Les tables conçues sont donc :

a) ARTICLES (code-doc, titre, article)

où :

code-doc : valeurs séquentielles (1 à 15) qu'identifie l'article

titre : contient le titre d'un article

article : contient le texte d'un article

Cette table est issue des faits 1, 3 et 4. Elle est en 3FN, parce que tous les attributs sont dépendants fonctionnellement de la clé et qu'il n'y a pas de dépendance transitoire entre les attributs non-clés.

b) SYNTAGMES (code du syntagme, syntagme)

où :

code du syntagme : code numérique séquentiel qu'identifie chaque syntagme

syntagme - contient le syntagme nominal proprement dit

Cette table est issue de la procédure de normalisation, étant donc en 3FN.

c) SYNTAGMES NIVEAU 1 (code 1, syntagme 1, nombre d'articles)

où :

code 1 : contient le code du syntagme nominal

syntagme 1 : contient le syntagme de niveau 1

nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.

Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 1.

Remarque : Cette table et les quatre autres suivantes sont redondantes par rapport à la table SYNTAGMES, mais elles sont toutes nécessaires car elles permettent une meilleure performance au sujet du temps d'accès étant donné que la séléction des syntagmes nominaux par niveau est faite avant la procédure de recherche. Elles sont toutes également issues du fait 7.

d) SYNTAGMES NIVEAU 2 (code 2, syntagme 2, nombre d'articles)

où :

code 2 : contient le code du syntagme nominal

syntagme 2 : contient le syntagme nominal de niveau 2

nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.

Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 2. Voir remarque dans la description de la rubrique 'c' plus haut.

e) SYNTAGMES NIVEAU 3 (code 3, syntagme 3, nombre d'articles)

où :

code 3 : contient le code du syntagme nominal

syntagme 3 : contient le syntagme nominal de niveau 3

nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.

Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 3. Voir remarque dans la description de la rubrique 'c' plus haut.

f) SYNTAGMES NIVEAU 4 (code 4, syntagme 4, nombre d'articles)

où :

code 4 : contient le code du syntagme nominal

syntagme 4 : contient le syntagme nominal de niveau 4

nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.

Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 4. Voir remarque dans la description de la rubrique 'c' plus haut.

g) SYNTAGMES NIVEAU 5 (code 5, syntagme 5, nombre d'articles)

où :

code 5 : contient le code du syntagme nominal

syntagme 5 : contient le syntagme nominal de niveau 5

nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.

Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 5. Voir remarque dans la description de la rubrique 'c' plus haut.

h) CENTRE DU SYNTAGME (code du centre, centre du syntagme)

où :

code du centre : code numérique séquentiel, identifie le centre du syntagme

centre du syntagme : contient le centre du syntagme nominal dans sa forme originale ; contient aussi les mots importants appelés dans ce travail de centre complementaire du syntagme nominal (voir chapitre III, section 3)

Cette table est issue du fait 12 et 14.

i) MOTS (code du centre, centre du syntagme)

où :

code du centre : code du centre de syntagme nominal dont la valeur est égale à celle de la table CENTRE du SYNTAGME

centre du syntagme : contient le mot résultant de la flexion en nombre d'un centre de syntagme nominal ou d'un mot qui équivaut à un centre de syntagme donné

Cette table est issue du fait 15.

j) REFERENCE RESUME (code, article)

où :

code : contient le code d'un syntagme nominal

article : contient le code d'identification de l'article

Cette table est conçue d'après le fait 5.

k) REFERENCE (code, article, paragraphe)

où :

code : contient le code du syntagme nominal

article : contient le code d'identification de l'article

paragraphe : contient le code du paragraphe d'un article donné

Cette table est issue des faits 2 et 6.

l) LIAISON CS - SN 1 (code syntagme niveau 1, code centre du syntagme)

où :

code syntagme niveau 1 : contient le code d'un syntagme nominal de niveau 1

code centre du syntagme : contient le code du centre de syntagme nominal

Cette table est issue des faits 12 et 13.

m) LIAISON SN 1 - SN 2 (code syntagme niveau 2, code syntagme niveau 1)

où :

code syntagme niveau 2 : contient le code du syntagme nominal niveau 2

code syntagme niveau 1 : contient le code du syntagme nominal niveau 1

Cette table est issue des faits 8, 9, 10 et 11.

n) LIAISON SN 2 - SN 3 (code syntagme niveau 3, code syntagme niveau 2)

où :

code syntagme niveau 3 : contient le code du syntagme nominal niveau 3

code syntagme niveau 2 : contient le code du syntagme nominal niveau 2

Cette table est issue des faits 8, 9, 10 et 11.

o) LIAISON SN 3 - SN 4 (code syntagme niveau 4, code syntagme niveau 3)

où :

code syntagme niveau 4 : contient le code du syntagme nominal niveau 4

code syntagme niveau 3 : contient le code du syntagme nominal niveau 3

Cette table est issue des faits 8, 9, 10 et 11.

p) LIAISON SN 4 - SN 5 (code syntagme niveau 5, code syntagme niveau 4)

où :

code syntagme niveau 5 : contient le code du syntagme nominal niveau 5

code syntagme niveau 4 : contient le code du syntagme nominal niveau 4

Cette table est issue des faits 8, 9, 10 et 11.

q) TABLE GROS INDEX (code du syntagme, syntagme, article, paragraphe, niveau, centre du syntagme, syntagme niveau inférieur)

où :

code du syntagme : est un code séquentiel numérique identifiant chaque occurrence des syntagmes nominaux, y inclus les syntagmes répétés.

syntagme : contient le syntagme nominal

article : contient le code de l'article d'où le syntagme a été extrait

paragraphe : contient le code du paragraphe d'où le syntagme a été extrait

niveau : contient le code du niveau du syntagme nominal

centre du syntagme : indique si le contenu de l'attribut du syntagme nominal de niveau inférieur est un centre du syntagme ou non

syntagme de niveau inférieur : contient le syntagme nominal de niveau inférieur par rapport à l'attribut du syntagme

La TABLE GROS INDEX est une table de travail d'où sont issues toutes les autres tables.

L'approche adoptée pour la construction de ce modèle de données a considéré l'utilisation du code identificateur d'un syntagme nominal au lieu d'utiliser le syntagme nominal lui-même comme étant la clé de chaque table. Cette option a été nécessaire en fonction de la limitation du logiciel Access en ce qui concerne la taille des champs d'une table, 256 caractères. En plus de ce problème, la comparaison entre deux champs numériques est plus performante que la comparaison entre deux champs textuels. Ce problème est encore plus important lorsqu'on a des syntagmes nominaux de niveau 4 et 5 qui atteignent parfois la limite de 256 caractères. Ainsi, pour réussir le développement de la maquette, on a choisi l'utilisation des codes des syntagmes nominaux au lieu des syntagmes eux-mêmes, ce qui explique la création d'un nombre plus grand de tables.

5. Structure de données : navigation dans l'arbre des syntagmes nominaux

La structure de données permettant la navigation dans l'arborescence des syntagmes nominaux étant déjà prête, il faut maintenant mettre en relation les tables de cette structure afin de construire effectivement la navigation sur l'arborescence.

a) La recherche des syntagmes nominaux de niveau un

Cette recherche est faite à partir d'une demande de l'utilisateur qui consiste à saisir un mot. Ensuite le système cherche dans la table MOTS si le mot demandé existe, et dans ce cas, prend le code du centre du syntagme nominal correspondant. Dès qu'il trouve le code du centre de syntagme nominal, le système trouve immédiatement dans la table LIAISON CS - NS 1 tous les codes des syntagmes de premier niveau associés à ce centre du syntagme. Les syntagmes nominaux vont être trouvés dans la table SYNTAGMES NIVEAU 1. De cette façon, on a conçue la relation suivante :

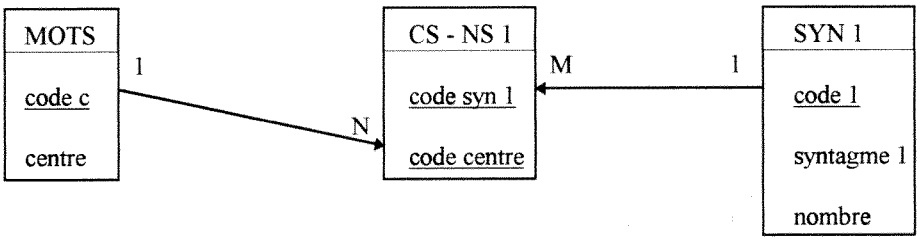


Figure 5. Relation pour la recherche des syntagmes de niveau un

Dans la figure 5 on a abrégé les noms des attributs pour mieux placer le dessin, le nom complet de chaque attribut se trouvant dans la section 4 de ce chapitre. Il est important de noter que l'illustration donne l'idée exacte de ce que l'on peut trouver pour chaque centre de syntagme plusieurs syntagmes de premier niveau. La figure montre aussi que l'inverse peut arriver, étant donné que l'on peut placer des mots alternatifs comme centre du syntagme lorsque ces mots possèdent autant d'importance que le centre du syntagme.

b) La recherche des syntagmes nominaux de niveau deux

La recherche des syntagmes nominaux de niveau deux est faite à partir du choix du syntagme nominal de premier niveau, demandé par l'utilisateur. Le système trouve le code du syntagme choisi dans la table SYNTAGMES NIVEAU 1 et puis cherche dans la table LIAISON SN 1 - SN 2 tous les syntagmes du deuxième niveau qui sont associés au

syntagme nominal de premier niveau choisi. Pour faire paraître les syntagmes associés, le système utilise la table SYNTAGMES NIVEAU 2 (voir figure 6).

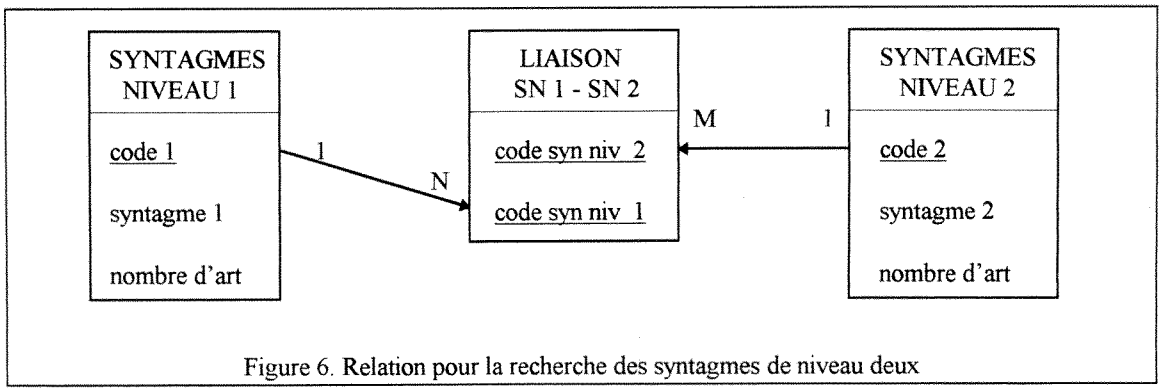


Figure 6. Relation pour la recherche des syntagmes de niveau deux

Ici on voit aussi que la structure LIAISON SN 1 - SN 2 permet qu'un syntagme quelconque de niveau un puisse être associé à plusieurs syntagmes de niveau deux et vice-versa.

c) La recherche des syntagmes de niveau trois

La démarche pour cette recherche est la même que pour les recherches précédentes. Il a suffit donc de changer les tables (voir la figure 7).

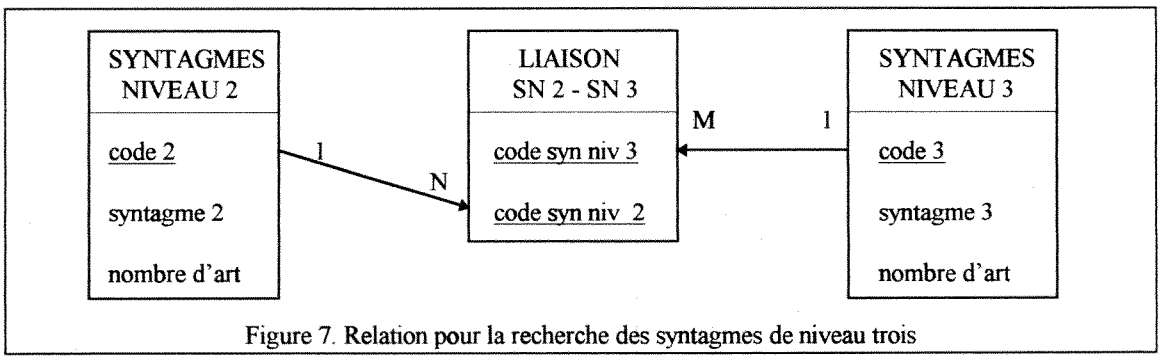


Figure 7. Relation pour la recherche des syntagmes de niveau trois

d) La recherche des syntagmes nominaux de niveau quatre et cinq

La recherche des syntagmes nominaux de niveau quatre et cinq ont la même démarche que toutes les autres déjà décrites (voir les figures 8 et 9 respectivement).

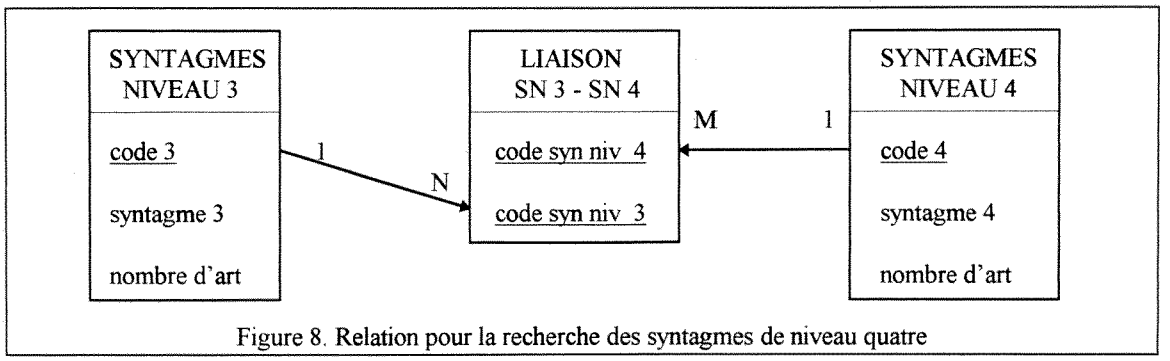


Figure 8. Relation pour la recherche des syntagmes de niveau quatre

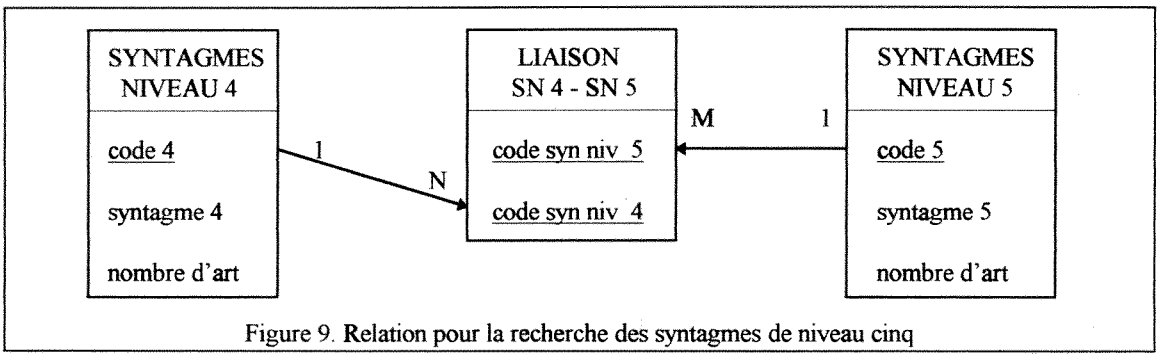


Figure 9. Relation pour la recherche des syntagmes de niveau cinq

f) La recherche des titres des articles à partir d'un syntagme nominal choisi

Pour chercher les titres qui correspondent aux articles d'où un syntagme nominal choisi a été extrait, on a conçu la relation suivante :

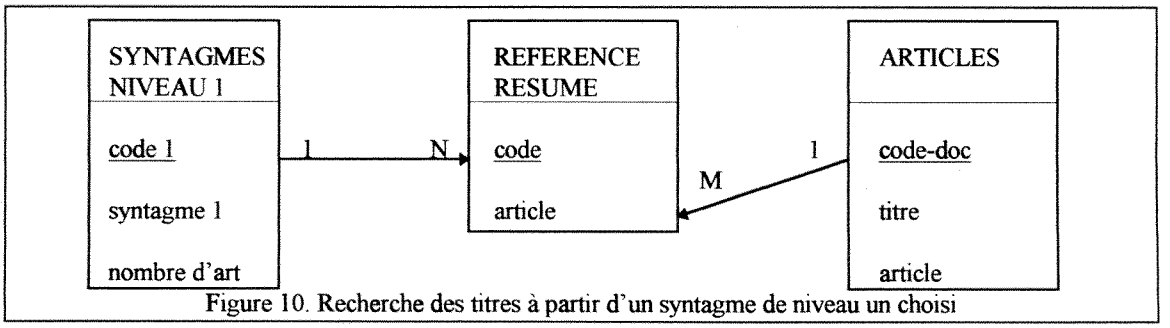


Figure 10. Recherche des titres à partir d'un syntagme de niveau un choisi

On voit dans la figure 10 que pour un syntagme nominal de niveau un, le système prend le code du syntagme nominal dans sa table et cherche dans la table REFERENCE RESUME tous les articles d'où ce syntagme a été extrait. A partir du code de chaque article, le système accède au titre respectif. Cette démarche est analogue pour les syntagmes nominaux d'un autre niveau, étant suffisant d'opérer la modification de la table des

SYNTAGMES NIVEAU 1 pour une autre table qui corresponde au niveau du syntagme souhaité.

6. Développement de la maquette du système de recherche d'information

Développer une maquette utilisant le logiciel Access signifie développer une application Access. Une telle application est composée d'objets d'utilisation directe comme les formulaires et les états, et d'objets d'utilisation indirecte comme les tables, les requêtes, les macros et les modules. Ces objets ont des propriétés paramétrables de façon à ce que l'on puisse donner l'aspect voulu à ces applications.

Par la suite on donnera une brève définition de ces objets, selon le manuel du logiciel.

> *« Un **formulaire** ... identifie les données à recueillir. Il permet d'enregistrer des informations dans une base de données, de les affranchir et les imprimer. »*[40]

> *« ... un **état** permet d'extraire et de présenter les données dans le format le mieux adapté à leur exploitation et à leur diffusion. Des étiquettes de publipostage, des factures, des rapports ... »*[41]

> *« Une **table** regroupe les données de même nature, ... chaque enregistrement d'une table contient des informations sur un élément en particulier,... les enregistrements d'une table sont constitués de champs. »*[42]

---

[40] MICROSOFT CO. *Microsoft Acces : Guide de l'utilisateur.* Ireland : 1994. p. 386

[41] Idem. p. 592

[42] Ibid. p. 136

*«Une **requête** sert à interroger des tables sur les données qu'elles contiennent. Sa structure indique précisément ... quelles données extraire. »*[43]

Pendant l'utilisation d'un formulaire, la modification des données d'un champ, l'action de cliquer sur un bouton ou le déplacement de la souris sont identifiés comme des événements, auxquels le logiciel Access peut réagir automatiquement. Pour personnaliser cette réaction du logiciel, on peut utiliser les macros et les procédures événementielles. Ces macros ou procédures sont exécutés lorsqu'un événement donné se produit. Pour cela, il faut attacher ces macros ou procédures à la propriété concernant l'événement. Un macro est composé d'un ensemble d'actions pré-définis comme OuvrirFormulaire, DéplacerDimensioner, CopierObjet etc. Une procédure est écrite en langage Access Basic, composé par des commandes et des fonctions (équivalents aux actions des macros).

Une fois que les caractéristiques du logiciel Access ont été connues et que la fonctionnalité a été comprise on s'est rendu compte que le développement de la maquette s'agissait uniquement de la création des **tables**, **requêtes** et **formulaires**. L'objet **état** n'a pas été utilisé parce qu'on n'a pas eu besoin de créer des fonctions d'impression de rapports.

6.1 Construction des tables

Les tables ont été construites selon le modèle de données décrit dans la section 4 de ce chapitre. Les rapports avec la définition de ces tables sont au Annexe E. Dans ces rapports toutes les caractéristiques de chaque table ont été présentées tels que : les noms, les formats et les tailles de ses champs. On a ainsi les tables suivantes :

a) Articles

Contient les codes, les titres et le contenu des articles.

---

[43] Ibid. p. 240

b) **Table des Syntagmes**

Contient tous les syntagmes nominaux indépendamment de leurs niveaux.

c) **Table des Mots**

Contient les codes des centres de syntagmes et les flexions en nombre de ces syntagmes.

d) **Table Référence**

Contient l'association entre les codes de chaque syntagme nominal, le code des articles et des paragraphes d'où ils ont été extraits.

e) **Table Référence Résumé**

Contient l'association des codes des syntagmes nominaux et les codes des articles d'où ces syntagmes ont été extraits.

f) **Table Centre du Syntagme**

Contient le code du centre de syntagme et les centres de syntagmes eux-mêmes.

g) **Table Syntagme Niveau 1**

Contient le code des syntagmes nominaux de premier niveau, les syntagmes nominaux eux-mêmes et le nombre d'articles où ces syntagmes apparaissent.

h) **Table Syntagme Niveau 2**

Contient le code des syntagmes nominaux de deuxième niveau, les syntagmes nominaux eux-mêmes et le nombre d'articles où ces syntagmes apparaissent.

i) **Table Syntagme Niveau 3**

Contient le code des syntagmes nominaux de troisième niveau, les syntagmes nominaux eux-mêmes et le nombre d'articles où ces syntagmes apparaissent.

j) **Table Syntagme Niveau 4**

Contient le code des syntagmes nominaux de quatrième niveau, les syntagmes nominaux eux-mêmes et le nombre d'articles où ces syntagmes apparaissent.

k) **Table Syntagme Niveau 5**

Contient le code des syntagmes nominaux de cinquième niveau, les syntagmes nominaux eux-mêmes et le nombre d'articles où ces syntagmes apparaissent.

l) **Table liaison CS - SN 1**

Contient l'association entre les codes des syntagmes de premier niveau et les codes des centres des syntagmes auxquels ils sont liés

m) **Table liaison SN 1 - SN 2**

Contient l'association entre les codes des syntagmes de deuxième niveau et les codes des syntagmes de premier niveau, auxquels ils sont liés.

n) **Table liaison SN 2 - SN 3**

Contient l'association entre les codes des syntagmes de troisième niveau et les codes des syntagmes de deuxième niveau auxquels ils sont liés.

o) **Table liaison SN 3 - SN 4**

Contient l'association entre les codes des syntagmes de quatrième niveau et les codes des syntagmes de troisième niveau auxquels il sont liés.

p) **Table liaison SN 4 - SN 5**

Contient l'association entre les codes des syntagmes de cinquième niveau et les codes des syntagmes de quatrième niveau auxquels ils sont liés.

q) **Table Gros Index**

Contient les champs de code du syntagme, syntagme, article, paragraphe, niveau, centre du syntagme, syntagme niveau inférieur.

6.2 Construction des requêtes

Pour la navigation dans l'arborescence des syntagmes nominaux on a construit des requêtes au moyen des relations présentées dans la section 5 de ce chapitre. Les rapports contenant la définition et les caractéristiques de chaque requête sont placés dans l'Annexe G.

Les requêtes construites sont les suivantes :

a) **Requête sur les SN 1**

Cette requête cherche, à partir d'un code de centre de syntagme, tous les syntagmes de niveau 1 associés.

b) **Requête sur les SN 2**

Cette requête cherche, à partir d'un code de syntagme de premier niveau, tous les syntagmes de niveau 2 associés.

c) **Requête sur les SN 3**

Cette requête cherche, à partir d'un code de syntagme de deuxième niveau, tous les syntagmes de niveau 3 associés.

d) **Requête sur les SN 4**

Cette requête cherche, à partir d'un code de syntagme de troisième niveau, tous les syntagmes de niveau 4 associés.

e) **Requête sur les SN 5**

Cette requête cherche, à partir d'un code de syntagme de quatrième niveau, tous les syntagmes du niveau 5 associés.

f) **Requête pour voir les titres 1**

Cette requête cherche, à partir d'un code de syntagme de premier niveau, tous les titres des articles d'où il a été extrait.

g) **Requête pour voir les titres 2**

Cette requête cherche, à partir d'un code de syntagme de deuxième niveau, tous les titres des articles d'où il a été extrait.

h) **Requête pour voir les titres 3**

Cette requête cherche, à partir d'un code de syntagme de troisième niveau, tous les titres des articles d'où il a été extrait.

i) **Requête pour voir les titres 4**

Cette requête cherche, à partir d'un code de syntagme de quatrième niveau, tous les titres des articles d'où il a été extrait.

j) **Requête pour voir les titres 5**

Cette requête cherche, à partir d'un code de syntagme de cinquième niveau, tous les titres des articles d'où il a été extrait.

6.3 Construction des formulaires

La construction des formulaires a été faite afin de procéder à l'interface entre l'utilisateur et l'application de la recherche d'information. Les rapports contenant les caractéristiques, les propriétés, les macros et les procédures événementielles pour chaque

formulaire développé apparaissent dans l'Annexe F. Ainsi, les formulaires construits, dans l'ordre de présentation, sont les suivants :

1. **Menu Général**

C'est le premier formulaire, dans l'ordre de présentation. Il a comme objectif de présenter les options de l'application pour que l'utilisateur puisse avoir le choix de la tâche à exécuter dans l'application. On a créé les boutons d'options suivants :

- construction de l'arborescence des syntagmes nominaux

Cette option permet de construire et d'ajuster l'arborescence des syntagmes nominaux. Elle permet aussi la création de nouveaux syntagmes nominaux. Une procédure événementielle associée a ce bouton ouvre le **Formulaire Saisir Syntagmes** ;

- recherche d'information

Ce bouton associé à une procédure événementielle ouvre le formulaire **Formreq**. À partir de ce formulaire l'utilisateur est guidé dans l'arborescence des syntagmes nominaux ;

- base de données

Cette option ouvre le module de création et de mise à jour de bases de données du logiciel Access. C'est une manière de mettre à jour la définition de la base de données et de l'application ;

- quitter l'application

ce bouton lorsqu'il est activé, permet de quitter l'application au moyen d'une procédure événementielle.

2. Formreq

Une fois l'option de recherche d'information choisie, l'application ouvre le formulaire **Formreq**. Le formulaire a comme objectif principal de recevoir la demande des utilisateurs au moyen d'un mot ou d'un centre de syntagme nominal. En réponse, l'application cherche les syntagmes nominaux de premier niveau associés a cette demande et les présente, s'ils s'y trouvent.

Ce formulaire a été développé en utilisant un sous-formulaire appelé **syntagme du premier niveau** ; ce sous-formulaire a comme fonction de montrer les syntagmes de premier niveau et la recherche de ces syntagmes est faite au moyen de la requête **Requête sur les SN 1**.

Une fois présentés tous les syntagmes nominaux de premier niveau, les utilisateurs peuvent choisir une des options suivantes :

- demander la recherche des syntagmes nominaux de deuxième niveau

Cette option peut être activée dès que l'utilisateur ait choisi un syntagme de premier niveau.

L'activation de cette option est faite à partir d'un clic sur le bouton correspondant à cette option ; une procédure événementielle est alors exécutée pour chercher les syntagmes et ouvrir le formulaire **voir syntagmes niveau deux** ;

- voir les titres d'où le syntagme de premier niveau choisi a été extrait

Une fois choisi un syntagme de premier niveau, l'utilisateur peut activer cette option ouvrant ainsi le formulaire **voir les articles 1** au moyen d'une procédure événementielle ;

- quitter l'option de recherche d'information

De même que dans tous les autres cas ci-dessous, une fois ce bouton activé, l'application exécute une procédure événementielle qui ferme le formulaire **Formreq** et passe le contrôle au formulaire **Menu Général**.

3. Voir syntagmes niveau deux

Ce formulaire est constitué d'un sous-formulaire, **syntagme du deuxième niveau**, qui a la fonction de présenter les syntagmes de niveau deux à partir du choix d'un syntagme de premier niveau et de la demande par l'utilisateur de rechercher ceux du niveau deux. Il utilise la requête **Requête sur les SN 2**.

Le formulaire a comme boutons d'option :

- chercher les syntagmes du troisième niveau

La procédure associée à ce bouton, cherche les syntagmes de niveau trois et ouvre le formulaire **voir syntagme niveau trois** ;

- voir les titres une fois choisi un syntagme de niveau deux

La procédure ouvre le formulaire **voir les articles 2** ;

- quitter le formulaire courant vers le formulaire **Formreq**

Si cette option est choisie, l'application ferme le formulaire courant ;

- faire une nouvelle recherche à partir d'un mot ou d'un centre de syntagme

Si cette option est choisie, l'application ferme le formulaire courant ;

- quitter l'application

Si cette option est choisie, l'application ferme le formulaire courant et le formulaire **Formreq**.

4. **Voir syntagmes niveau trois**

Ce formulaire est constitué d'un sous-formulaire, **syntagme du troisième niveau**, qui a la fonction de présenter les syntagmes de niveau trois à partir du choix d'un syntagme de deuxième niveau et de la demande par l'utilisateur de rechercher ceux du niveau trois. Il utilise la requête **Requête sur les SN 3**.

Le formulaire a comme boutons d'option :

- chercher les syntagmes de quatrième niveau ;

La procédure associée à ce bouton, cherche les syntagmes de niveau quatre et ouvre le formulaire **voir syntagme niveau quatre** ;

- voir les titres une fois choisi un syntagme de niveau trois

La procédure ouvre le formulaire **voir les articles 3** ;

- quitter le formulaire courant vers le formulaire **voir syntagme niveau deux**

Si cette option est choisie, l'application ferme le formulaire courant ;

- faire une nouvelle recherche à partir d'un mot ou d'un centre de syntagme

Si cette option est choisie, l'application ferme le formulaire courant et le formulaire **voir syntagmes niveau deux** ;

- quitter l'application

Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **voir syntagmes niveau deux** et le formulaire **Formreq**.

5. Voir syntagmes niveau quatre

Ce formulaire est constitué d'un sous-formulaire, **syntagmes du quatrième niveau**, qui a la fonction de présenter les syntagmes de niveau quatre à partir du choix d'un syntagme de troisième niveau et de la demande par l'utilisateur de rechercher ceux de niveau quatre. Il utilise la requête **Requête sur les SN 4**.

Le formulaire a comme boutons d'option :

- chercher les syntagmes de cinquième niveau

La procédure associée à ce bouton, cherche les syntagmes de niveau cinq et ouvre le formulaire **voir syntagmes niveau cinq** ;

- voir les titres une fois choisi un syntagme de niveau quatre

La procédure ouvre le formulaire **voir les articles 4** ;

- quitter le formulaire courant vers le formulaire **voir syntagmes niveau trois**

Si cette option est choisie, l'application ferme le formulaire courant ;

- faire une nouvelle recherche à partir d'un mot ou d'un centre de syntagme

Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **voir syntagmes niveau trois** et le formulaire **voir syntagmes niveau deux** ;

- quitter l'application

Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **voir syntagmes niveau trois**, le formulaire **voir syntagme niveau deux** et le formulaire **Formreq**.

6. Voir syntagmes niveau cinq

Ce formulaire est constitué d'un sous-formulaire, **syntagmes du cinquième niveau**, qui a la fonction de présenter les syntagmes de niveau cinq à partir du choix d'un syntagme de quatrième niveau et de la demande par l'utilisateur de rechercher ceux de niveau cinq. Il utilise la requête **Requête sur les SN 5**.

Le formulaire a comme boutons d'option :

- voir les titres une fois choisi un syntagme de niveau cinq ;

La procédure ouvre le formulaire **voir les articles 5** ;

- quitter le formulaire courant vers le formulaire **voir syntagmes niveau quatre**

Si cette option est choisie, l'application ferme le formulaire courant ;

- faire une nouvelle recherche à partir d'un mot ou d'un centre de syntagme

Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **voir syntagmes niveau quatre**, le formulaire **voir syntagmes niveau trois** et le formulaire **voir syntagmes niveau deux** ;

- quitter l'application

Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **voir syntagmes niveau quatre,** le formulaire **voir syntagmes niveau trois**, le formulaire **voir syntagmes niveau deux** et le formulaire **Formreq**.

7. **Voir les articles 1, voir les articles 2, voir les articles 3, voir les articles 4 et voir les articles 5**

Ces formulaires utilisent des sous-formulaires pour montrer les titres des articles ; ces sous-formulaires sont respectivement : **Montre Articles 1**, **Montre Articles 2**, **Montre Articles 3**, **Montre Articles 4 et Montre Articles 5**.

Les sous-formulaires ont la fonction de rechercher les titres des articles et de les présenter dans le formulaire correspondant.

Les requêtes responsables pour ces recherches sont respectivement pour chaque sous-formulaire : **Requête pour voir les titres 1**, **Requête pour voir les titres 2**, **Requête pour voir les titres 3**, **Requête pour voir les titres 4** et **Requête pour voir les titres 5**.

Ces formulaires ont comme boutons d'option :

- voir l'article selon le choix du titre par l'utilisateur

La procédure ouvre le formulaire **Montre Doc** ;

- quitter le formulaire vers le dernier formulaire ouvert

L'application exécute la procédure de fermeture du formulaire courant.

8. **Montre Doc**

La source pour ce formulaire est la table **Articles**. Ce formulaire a comme objectif de montrer l'article selon le choix du titre par l'utilisateur.

9. **Formulaire Saisir Syntagmes**

Les objectifs de ce formulaire sont : réviser les syntagmes extraits du corpus, construire l'arborescence des syntagmes nominaux et inclure des nouveaux syntagmes nominaux.

La source pour ce formulaire est la **Table Gros Index.**

L'utilisation de ce formulaire a été très importante pour le chargement de la base de données et la construction de l'arborescence des syntagmes nominaux, parce qu'il a permis de faire : a) la révision des syntagmes nominaux extraits ; b) l'association entre les syntagmes ; c) l'attribution de centres de syntagmes à chaque syntagme nominal de premier niveau ; et d) l'attribution de niveaux aux syntagmes en observant le caractère relatif d'association entre deux syntagmes donnés.

7. Conclusion

Les procédures de développement et d'exploitation de la maquette ont consisté de deux étapes : la première a été une étape d'expérimentation où une petite maquette a été construite avec un modèle de données similaire au modèle final présenté dans ce chapitre. Pendant cette phase : la mise en relation entre les syntagmes nominaux a été faite directement entre les suites de caractères des syntagmes nominaux et non pas par des codes d'identification ; la mise en arbre des syntagmes nominaux a considéré les niveaux absolus des syntagmes et les syntagmes de premier niveau ont été mis en relation direct avec les centres des syntagmes. On a chargé cinq articles dans la base de données pour tester cette maquette. Cette étape était importante pour connaître les limitations du logiciel Access et du modèle d'arborescence initial. Les problèmes trouvés dans cette phase et les solutions adoptées pour construire la maquette définitive vont être présentés dans le chapitre III.

Dans la deuxième étape on s'est occupé de la construction de la maquette définitive tenant en compte les limitations du logiciel et les solutions trouvées à propos de l'arborescence des syntagmes nominaux. Le développement présenté dans ce chapitre concerne à la maquette définitive.

Cette démarche a été très utilie dans le processus d'apprentissage du logiciel Access et étant donné la limitation du temps, on a choisi d'apprendre pendant qu'on l'utilisait.

« Personne ne devient jamais maître dans un domaine où il n'a pas connu l'impuissance, et qui souscrit à cela saura aussi que cette impuissance ne se trouve ni au début ni avant l'effort entrepris, mais en son centre »

Walter BENJAMIN (1892-1940)

Chapitre III

Exploitation de la maquette

A partir de la démarche utilisée pour construire la maquette du système de recherche d'information, on présentera ici les remarques relevées en exploitant la première maquette; cela a permis de connaître les limitations du logiciel Access et les problèmes relatifs à la construction de l'arborescence ainsi que la définition des centres des syntagmes nominaux de premier niveau. On discutera aussi les solutions possibles et lesquelles ont été adoptées.

1. Chargement de la base de données dans la maquette

Pour la première étape de la construction et de l'exploitation de la maquette du système de recherche d'information, on n'a chargé que 5 articles dans la base de données.

La procédure de changement des cinq premiers articles dans la base de données a été faite comme suit :

1. Identification des niveaux des syntagmes nominaux dans les fichiers Word 6.0a ;

2. Importation des syntagmes nominaux, en format Word 6.0a dans la table de travail GROSIND ;

3. Création de la table GROS INDEX à partir de la table GROSIND ;

4. Détermination des centres des syntagmes nominaux pour chaque syntagme de premier niveau et création de la table de Centre du Syntagme ;

5. Création des tables des syntagmes, tables des syntagmes niveau 1, 2, 3, 4 et 5 à partir de requêtes de sélection sur la table Gros Index ;

6. Construction de l'arborescence des syntagmes nominaux à l'aide de la création des tables de liaison entre les syntagmes nominaux d'un niveau donné avec son correspondant de niveau inférieur (syntagmes nominaux niveau 2 avec syntagmes nominaux niveau 1, syntagmes nominaux niveau 3 avec les correspondants niveau 2 et ainsi de suite). Pour établir cette liaison on a créé un formulaire pour chaque association ;

7. Création des tables de références et des articles ;

8. Comptage du nombre d'articles d'où chaque syntagme a été extrait. Ce comptage a été fait en utilisant des requêtes de sélection et des requêtes d'ajout.

Le travail de saisir les cinq premiers articles dans la base de données et de construire l'arborescence des syntagmes nominaux a été très lourd, étant donné que toute la procédure était manuel et que la construction de chaque niveau d'arborescence prenait en compte un syntagme à la fois et à chaque niveau.

A partir de cette expérience, pour le chargement définitif du corpus dans la base de données, on a adopté les procédures suivantes :

1. Importation des fichiers Word 6.0a, contenant les syntagmes nominaux, groupés par chaque article, dans la table de GROSIND (voir annexe E) ;

2. Création, à partir de la table GROSIND, de la table GROS INDEX, mettant le champ des syntagmes nominaux, tant dans le champ syntagme que dans le champ syntagme nominal inférieur ; cette procédure a évité la tâche de saisir manuellement chaque syntagme nominal de niveau inférieur ;

3. Révision, à l'aide du formulaire Saisir Syntagmes, de tous les syntagmes nominaux et changement du champ syntagmes nominaux inférieur - étant donné que ce champ a été crée à l'image du champ syntagme comme on a souligné ci-dessus. Dans cette révision, on a définit aussi le niveau relatif d'association entre les syntagmes nominaux. Le développement de ce formulaire a permis de rendre la tâche de construction de l'arborescence et de définition des centres des syntagmes nominaux moins lourde que dans l'expérimentation initiale ;

4. Introduction des flexions en nombre des centres de syntagme nominal au moyen du formulaire X AJUSTE CENTRE dans la table X TABLE CENTRE DU SYNTAGME, qui est à l'origine de la TABLE DES MOTS ;

5. Création de toutes les tables définies dans la maquette, au moyen des requêtes de sélection et d'ajout, à partir de la TABLE GROS INDEX. Dans l'annexe G on présente la définition de ces requêtes dans la séquence de création des tables.

L'expérimentation de la maquette avec les cinq premiers articles a permis d'observer les limitations suivantes au sujet du logiciel : a) la taille maximale d'un champ type texte est de 256 caractères; b) le logiciel n'arrive pas à bien travailler avec une requête d'ajout dont la somme de la taille des champs soit plus grande que 256 caractères ; c) la recherche de champ type texte est plus lente que n'importe quel type de champ. Parmi ces limitations la plus importante est la limitation du nombre de caractères (256) empêchant la liaison de deux champs ou plus, puisque ce type d'opération est plus communne dans une

procédure de recherche d'information lorsqu'on a pour base l'arborescence des syntagmes nominaux. Pour éviter ces problèmes, dans la maquette finale, on a créé un code unique pour chaque syntagme nominal. Ainsi toutes les opérations de comparaison et d'ajout sont faites sur le code et non pas sur le texte du syntagme nominal. Ainsi pour éviter d'atteindre la limite de la longueur d'un champ on a décidé de définir la taille de chaque champ comme étant de 150 caractères.

Comme conséquence de cette limitation on n'a eu que deux solutions pour stocker les textes des articles. Une solution étant de les considérer comme un objet importé, la deuxième de les mettre dans un champ type mémo. Aucune de ces deux solutions n'était pas la bonne, car elles ne permettaient pas de traiter les textes. Pour la maquette il fallait avoir des possibilités de distinction des syntagmes nominaux dans les textes lorsqu'on demande à voir les articles. Ainsi, parmi les deux solutions la seconde étant la moins contraignante, on a gardé les textes des articles dans les champs type mémo. Cela a permis de présenter l'article en entier, ce qui autrement serait impossible.

Les problèmes relatifs au comportement des syntagmes nominaux dans son organisation en arbre et aux centres des syntagmes seront discutés dans la section suivante.

2. Comportement des syntagmes nominaux dans l'arborescence

L'approche initiale pour la construction de l'arborescence a été établie en considérant que les syntagmes nominaux étaient d'un niveau absolu les uns par rapport aux autres. Ainsi, on a créé des tables qu'associent un syntagme nominal d'un niveau donné avec le syntagme nominal d'un rang immédiatement inférieur.

Par exemple, dans la table LIAISON SN 1 - SN 2 on a associé pour chaque syntagme nominal de deuxième niveau tous les syntagmes de premier niveau d'où ils ont été extraits. Pour la table LIAISON SN 2 - SN 3, on a associé pour chaque syntagme de troisième niveau tous les syntagmes de deuxième niveau d'où ils ont été extraits et ainsi de

suite. D'une manière générale, on peut représenter graphiquement cette arborescence comme dans la figure 11.

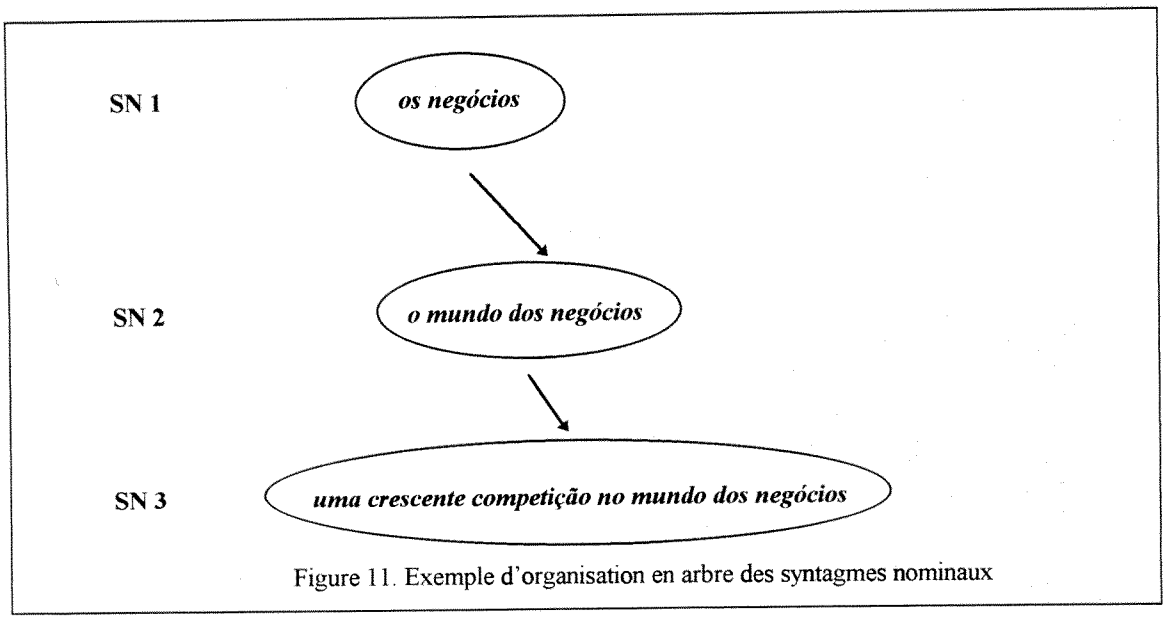


Figure 11. Exemple d'organisation en arbre des syntagmes nominaux

Toutefois, on n'a pas toujours trouvé des syntagmes dont les niveaux inférieurs soient directement associés au niveau consécutivement supérieur. On trouve parfois des syntagmes nominaux dont les différents niveaux se disposent indépendamment. Ce sont des syntagmes nominaux avec double rection.

Dans l'exemple de la figure 12, l'approche adoptée initialement n'a pas fonctionné car on n'a pas pu arriver au syntagme de plus haut niveau au moyen du syntagme de premier niveau ***a informação estratégica***.

Pour résoudre ce problème, au lieu d'organiser l'arborescence en considérant les niveaux de chaque syntagme nominal de manière absolue, on a construit l'arborescence en gardant les niveaux relatifs d'association entre les syntagmes nominaux.

Dans la figure 13 on montre l'arborescence en considérant les niveaux relatifs d'associations entre les syntagmes nominaux. Le syntagme nominal *a aceitação da informação estratégica na definição do futuro da empresa* est, dans cette nouvelle approche, à la fois syntagme nominal de deuxième niveau par rapport au syntagme nominal

a informação estratégica et syntagme nominal de quatrième niveau par rapport au syntagme nominal *a empresa*.

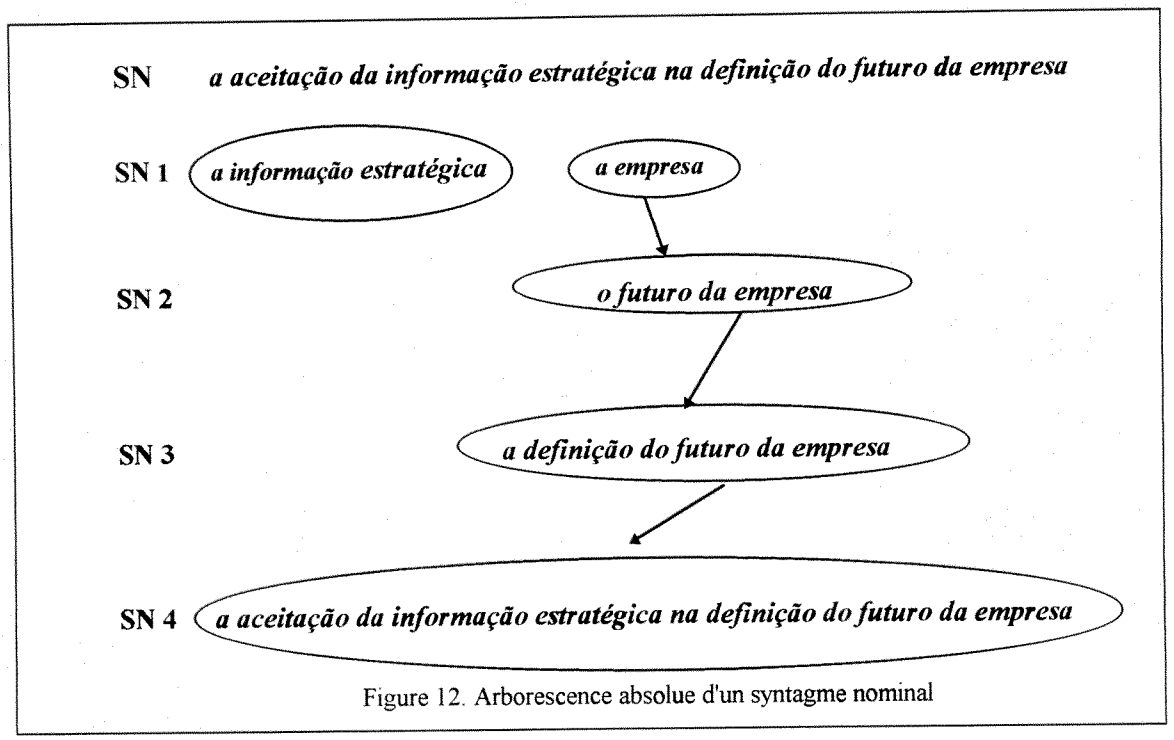


Figure 12. Arborescence absolue d'un syntagme nominal

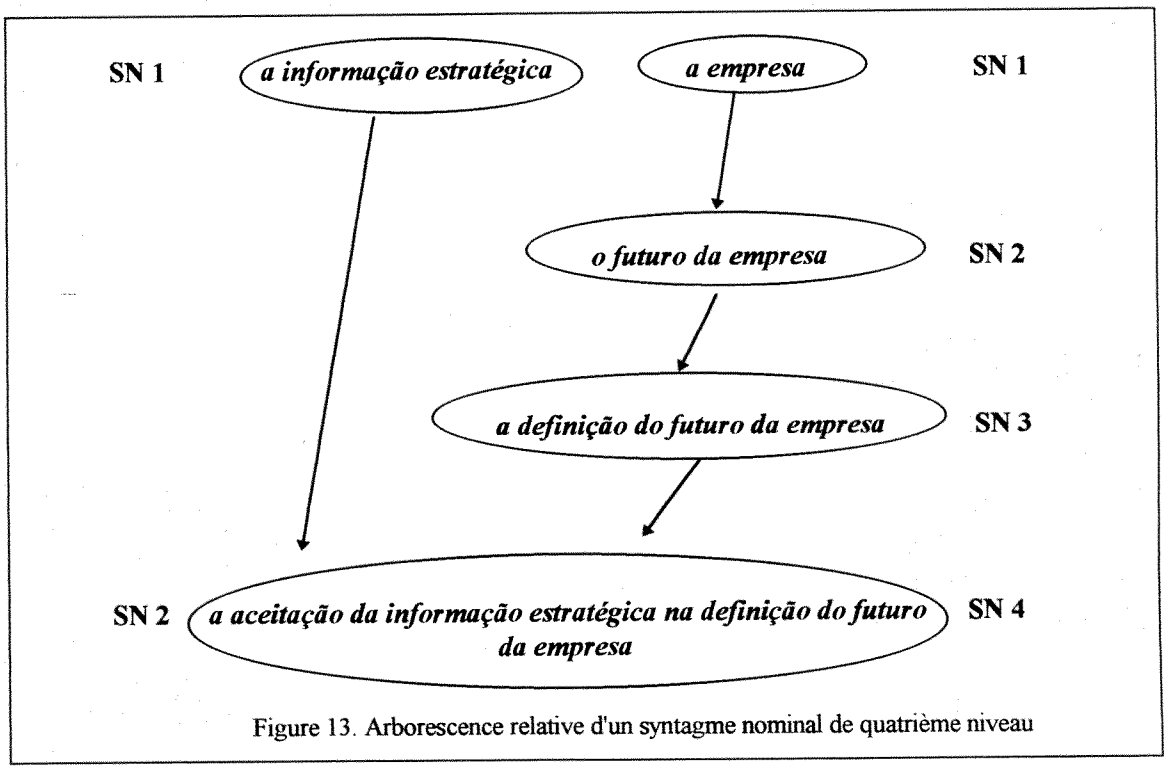


Figure 13. Arborescence relative d'un syntagme nominal de quatrième niveau

Ce changement n'altère en rien la maquette du système de recherche d'information, étant donné que la construction de l'arborescence des syntagmes nominaux est faite manuellement.

3. Centres complémentaires des syntagmes nominaux

Lorsqu'une recherche d'information était faite dans la première maquette, cherchant toujours l'information à partir des centres des syntagmes nominaux, on a constaté l'existence de silence (nombre de références ou documents pertinents manqués à la suite d'une recherche d'information, alors qu'ils existent dans la base de données) par rapport aux mots qui sont parfois des centres des syntagmes nominaux alors que dans d'autres situations ils ne le sont pas. Usuellement ces mots apparaissent dans les syntagmes nominaux qui possèdent une expansion prépositionnelle, comme par exemple : *Os sistemas de informação científico-tecnológico*.

Le centre du syntagme nominal est : *sistemas*. Or, bien que le mot *informação*, dans ce cas, n'est pas le centre du syntagme, il est quand même important pour la recherche d'information. Lorsqu'on fait la recherche à partir du centre du syntagme nominal *informação*, on ne trouve pas les documents indexés par le syntagme nominal *os sistemas de informação científico-tecnológico*. Cela produit du silence. Pour résoudre ce problème on propose la création de la figure du centre complémentaire des syntagmes nominaux. Ce sont des mots qui ont une importance égale aux centres des syntagmes nominaux.

Du point de vue de la maquette, il faut créer une structure capable de permettre la recherche non seulement à partir des centres des syntagmes nominaux, mais aussi à partir des centres complémentaires des syntagmes nominaux. Pour cela, il y a deux solutions possibles : a) créer une table de mots complémentaires composés par les mots qui ne sont pas des centres des syntagmes nominaux, mais qui sont quand même très important pour la recherche ; b) inclure ces mots dans la TABLE CENTRE DU SYNTAGME, bien qu'ils ne le soient pas.

La solution 'a' est plus intéressant du fait que la TABLE CENTRE DU SYNTAGME resterait intègre. Or, ce type de solution est cependant la moins performante en considérant que le système doit faire la recherche dans deux tables au lieu de la faire dans une seule.

La solution 'b' qui est moins intéressante du point de vue de la structure de données, montre qu'on pourra avoir des mots dans la TABLE CENTRE DU SYNTAGME qui ne sont pas vraiment des centres des syntagmes nominaux. Par contre, du point de vue de la performance du système de recherche d'information, c'est la solution la plus indiquée, car le système ne fera la recherche que dans une seule table.

4. Centres des syntagmes nominaux et ses flexions

Dans la première version du système de recherche d'information on a laissé les centres des syntagmes nominaux tels qu'ils sont apparus dans les syntagmes et dans le corpus. On s'est aperçu, en exploitant le système, que les résultats étaient faibles lorsqu'on essayait de chercher des syntagmes nominaux dont le centre était *informação*. Le système trouvait *a informação, a informação científica, a informação técnica, etc.* Or, le système ne trouvait pas des syntagmes comme : *as informações, as informações científicas, as informações industriais, as informações organizacionais, as informações técnicas, etc...* On avait encore du silence !

Pour résoudre ce problème on a considéré deux solutions possibles : 1) la mise en oeuvre des opérateurs de troncature ; et 2) le traitement des flexions, en nombre et en genre, des centres des syntagmes nominaux en créant une table de mots qui soit équivalente aux centres des syntagmes nominaux.

La solution de mise en oeuvre des opérateurs de troncature n'est pas une bonne solution car on s'est trouvé devant la possibilité d'existence des mots dont la flexion n'est

pas fait à la fin. Ainsi, cette alternative ne peut résoudre que partialement le problème et la solution utilisée a été la deuxième alternative.

Quoi qu'il en soit, dans un système d'information, on trouve important l'adoption des deux solutions, puisqu'elles rendront le système plus souple et dont les résultats auront moins de silence. Dans cette étape de la recherche d'information, l'augmentation du bruit (proportion de références ou documents non-pertinents trouvés à l'aide d'une recherche d'information) est plus intéressante et souhaitable plutôt qu'un taux élevé de silence. Cela se justifie parce que l'usager, en ce moment de la recherche, est en train de choisir les syntagmes qu'appartiennent encore au premier niveau et cela suppose un plus grand choix. On voit ici l'importance de l'interaction entre le système de recherche d'information et l'utilisateur.

Après l'implémentation de la solution choisie et lorsque on a exploité la maquette en vérifiant la navigation dans l'arborescence des syntagmes nominaux, on a observé l'apparition de plusieurs syntagmes nominaux de même signification mais ayant un lexique différent. Exemple : à partir d'une demande utilisant le mot *informação* en tant que centre de syntagme nominal, on trouve des syntagmes nominaux comme *a informação, as informações, informações* parmi d'autres syntagmes nominaux associés à ce centre de syntagme nominal. Si l'on veut connaître tous les syntagmes nominaux de deuxième niveau associés au syntagme nominal *a informação*, il faut chercher aussi les syntagmes nominaux associés à *as informações* et *informações*, étant donné qu'ils font référence au même sujet. Cependant la maquette n'offre pas cette possibilité.

Une solution possible, serait de développer une simulation de l'opérateur booléen « ou », étant donné que l'interface développé est orientée par menus.

5. Statistique descriptive sur les syntagmes nominaux

Les associations et comportements des syntagmes nominaux dans l'arborescence peuvent être vus au moyen des statistiques descriptives. La figure 14, montre la distribution des syntagmes nominaux dans les quinze (15) articles avec toutes ses occurrences.

Article	Niveau 1	Niveau 2	Niveau 3	Niveau 4	Niveau 5	Total
1	378	221	53	22	7	681
2	357	161	51	6	1	576
3	338	152	50	15		555
4	290	153	47	9	3	502
5	520	136	47	23		726
6	315	178	87	8		588
7	368	175	61	20		624
8	311	210	90	22	9	642
9	288	200	76	21	5	590
10	266	150	50	13		479
11	267	149	41	10		467
12	362	194	46	9		611
13	278	168	66	6	1	519
14	396	209	53	13	1	672
15	352	165	54	13	2	586
Total	5086	2621	872	210	29	8818

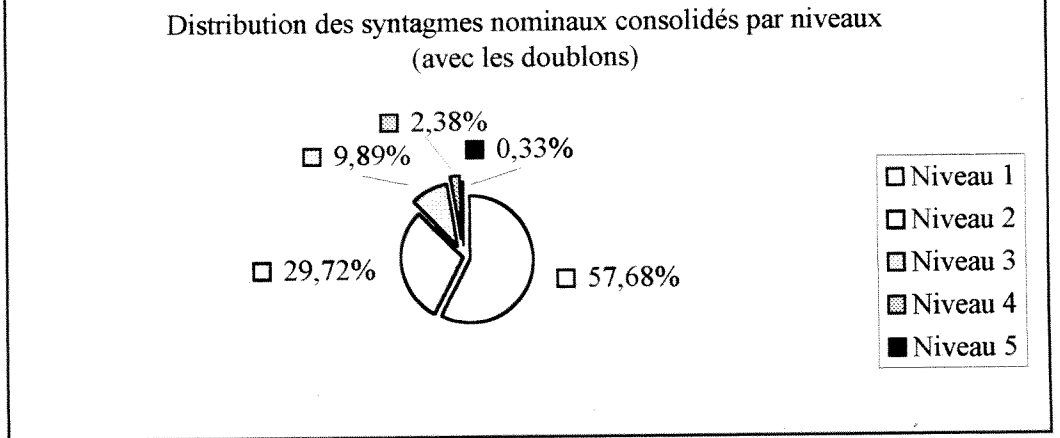


Figure 14. Distribution des syntagmes nominaux consolidés par niveaux (avec les doublons)

Il faut dire que l'occurrence multiple d'un même syntagme nominal résulte non seulement de l'occurrence naturel dans les articles, mais du calcul des anaphores et des syntagmes nominaux avec factorisation. Ces calculs ont produit plusieurs syntagmes

nominaux. C'est pour cela qu'on trouve parfois la répétition des syntagmes nominaux dans un même paragraphe.

La figure suivante, figure 15, montre la distribution des syntagmes nominaux dans le corpus, sans les doublons. Il faut dire que, bien que dans ce tableau les doublons des syntagmes nominaux dans un même article n'apparaissent pas, on peut trouver encore des doublons des syntagmes nominaux parmi l'ensemble des syntagmes nominaux de deux articles ou plus.

Article	Niveau 1	Niveau 2	Niveau 3	Niveau 4	Niveau 5	Total
1	238	185	49	21	7	500
2	284	142	51	6	1	484
3	177	125	45	15		362
4	154	131	47	9	3	344
5	396	124	47	23		590
6	170	139	86	8		403
7	259	139	57	20		475
8	199	175	79	22	9	484
9	197	179	71	19	5	471
10	179	113	48	13		353
11	165	119	40	10		334
12	274	176	43	9		502
13	198	139	62	6	1	406
14	266	168	51	13	1	499
15	233	130	50	13	2	428
Total	3389	2184	826	207	29	6635

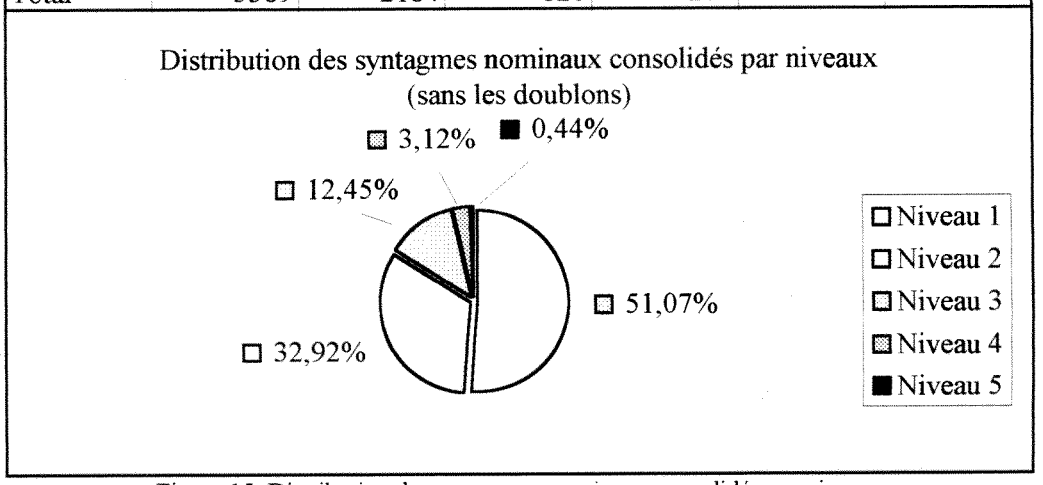


Figure 15. Distribution des syntagmes nominaux consolidés par niveaux (sans les doublons dans chaque article)

La comparaison de ces deux tableaux (figures 14 et 15), indique une chute, en pourcentage, des syntagmes nominaux de niveau un et une augmentation, aussi en pourcentage, des syntagmes nominaux des autres niveaux. Cette augmentation s'explique en fonction de ce qu'il y a eu une quantité plus grande de doublons des syntagmes nominaux de niveau un par rapport à ceux des autres niveaux. On voit que les doublons des syntagmes nominaux de niveaux plus élevés (3, 4 et 5) sont plus rares que ceux de niveaux moins élevés (1 et 2). Le pourcentage de doublons des syntagmes nominaux, dans chaque niveau a été : a) niveau un, 50,07% ; b) niveau deux, de 20,01% ; c) niveau trois, de 5,57% ; d) niveau quatre, de 1,45% ; et e) niveau cinq, de 0%.

Depuis la construction de la base de données, où on a créé des tables pour chaque niveau des syntagmes nominaux, il a été possible de connaître sa distribution finale sans doublons, c'est-à-dire le nombre des syntagmes nominaux uniques pour chaque niveau, selon la figure 16.

	Syntagme niveau 1	Syntagme niveau 2	Syntagme niveau 3	Syntagme niveau 4	Syntagme niveau 5	Total
N°. de syntagmes	2928	2161	826	207	29	6151

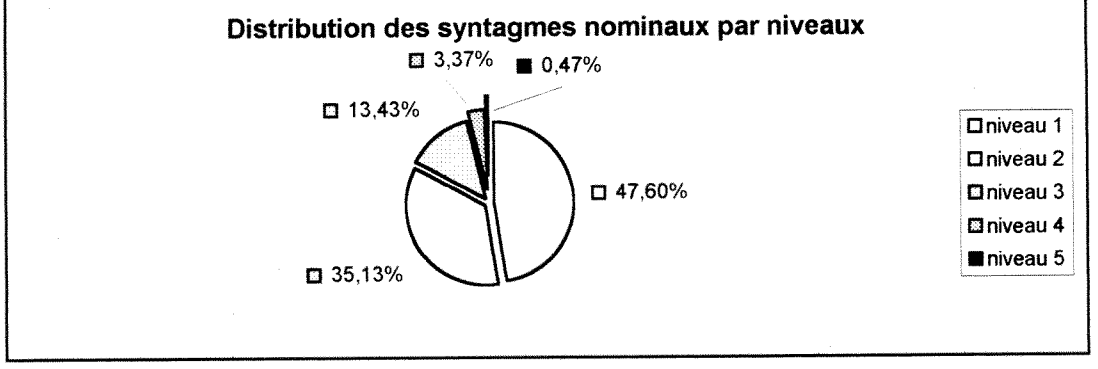


Figure 16. Distribution des syntagmes nominaux consolidés par niveaux (sans doublons)

Comme dans la comparaison entre le tableau Distribution des syntagmes nominaux avec doublons et le tableau Distribution des syntagmes nominaux sans doublons dans chaque article, la figure 16 indique un petit accroissement dans le pourcentage des syntagmes nominaux des niveaux 2, 3, 4 et 5, en opposition à la chute du pourcentage des syntagmes nominaux de premier niveau. Dans les deux cas, la quantité de doublons des syntagmes nominaux de premier niveau est plus grande par rapport à ceux des autres

niveaux. D'ailleurs, il n'y a pas eu des doublons sur les syntagmes nominaux des niveaux 3, 4 et 5.

La constatation de l'inexistence des doublons des syntagmes nominaux, à partir des niveaux 3, 4 et 5, entre les articles du corpus est cohérente avec l'idée selon laquelle ces niveaux sont responsables du raffinement de la recherche d'information.

Dans l'arborescence on trouve deux genres de syntagmes nominaux ; le premier qui n'est pas associé à aucun syntagme nominal, et qu'on les appellera désormais syntagmes nominaux terminaux, et le deuxième qui se trouve associé aux syntagmes nominaux de niveau supérieur, auxquels on appellera syntagmes intermédiaires. Les figures 17 et 18 montre la quantité des syntagmes nominaux terminaux et celle des intermédiaires.

	NIVEAU 1		NIVEAU 2		NIVEAU 3		NIVEAU 4		TOTAL	
	N°.Syn.	%	N°.Syn.	%	N°.Syn.	%	N°.Syn.	%	N°.Syn.	%
Synt. Terminaux	1808	61,75%	1512	69,97%	638	77,24%	179	86,47%	4137	67,58%
Syn. Intermédiaires	1120	38,25%	649	30,03%	188	22,76%	28	13,53%	1985	32,42%
Total	2928	100,00%	2161	100,00%	826	100,00%	207	100,00%	6122	100,00%

Figure 17. Syntagmes nominaux terminaux versus syntagmes nominaux intermédiaires

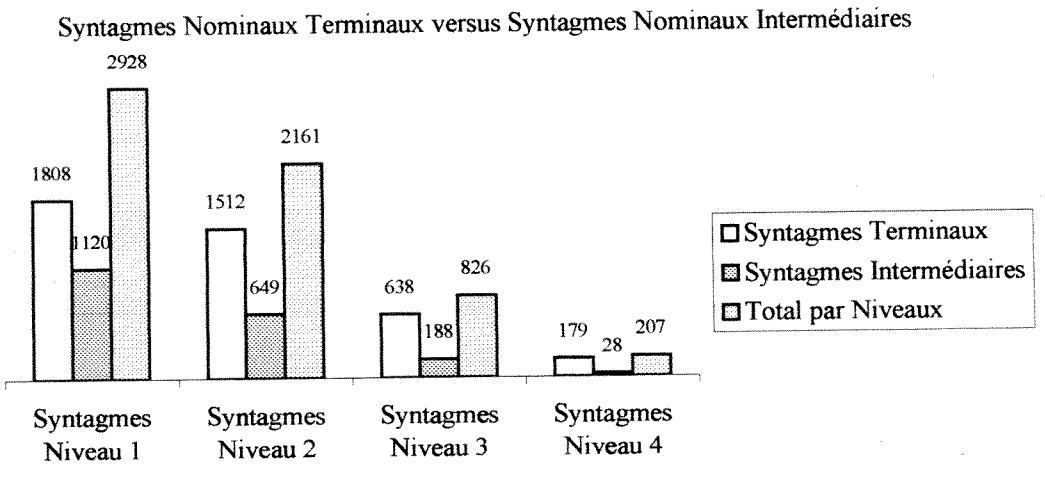


Figure 18. Histogramme des syntagmes nominaux terminaux et syntagmes nominaux intermédiaires par niveaux

Ainsi, on a 1808 syntagmes nominaux de premier niveau que sont aussi terminaux parce qu'ils ne sont pas associés à aucun syntagme nominal de niveau deux. Comme syntagmes nominaux intermédiaires de premier niveau on a 1120 syntagmes nominaux

associés aux syntagmes nominaux de deuxième niveau. Par rapport aux syntagmes nominaux de niveau deux, 1512 sont terminaux et 649 sont associés aux syntagmes nominaux de troisième niveau. La même interprétation est donnée aux syntagmes nominaux de niveau trois et quatre. On se rend compte qu'il y a une décroissance d'environ 8% sur le nombre des syntagmes nominaux d'un niveau donné par rapport à ceux d'un niveau immédiatement supérieur. Ce fait démontre la capacité de raffinement que l'arborescence des syntagmes nominaux possède dans la procédure de navigation et de recherche d'information.

Par rapport à l'arborescence des syntagmes nominaux on a construit le tableau de la figure 19.

	CS - NS 1	SN 1 - SN 2	SN 2 - SN 3	SN 3 - SN 4	SN 4 - SN 5	Total
N°. Associations	3150	2402	852	207	29	6640
N° Relatif d'As.	47,44%	36,17%	12,83%	3,12%	0,44%	100,00%

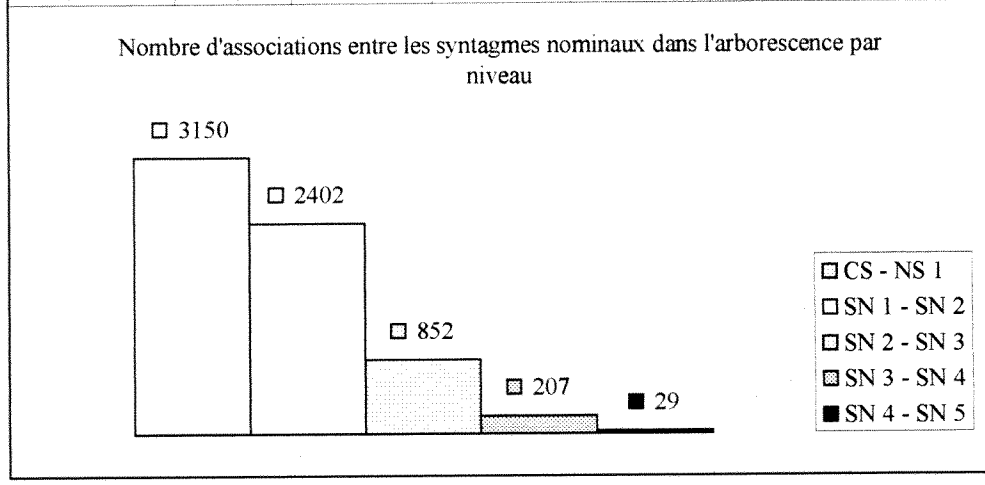


Figure 19. Nombre d'associations entre les syntagmes nominaux dans l'arborescence par niveau

La figure 19 montre qu'il y a 3150 associations entre les centres de syntagmes nominaux et les syntagmes nominaux de premier niveau, 2402 associations entre les syntagmes nominaux intermédiaires de premier niveau et les syntagmes nominaux de deuxième niveau, 852 associations entre les syntagmes nominaux intermédiaires de deuxième niveau et les syntagmes nominaux de troisième niveau, et ainsi de suite.

L'analyse de ce tableau montre la capacité de raffinement que l'arborescence des syntagmes nominaux permet dans une procédure de navigation et de recherche d'information. Une autre constatation vient du fait que dans les associations entre les centres de syntagme nominal et les syntagmes nominaux de premier niveau aussi bien qu'entre les syntagmes nominaux intermédiaires de premier niveau et les syntagmes nominaux de deuxième niveau et entre les syntagmes nominaux intermédiaires du deuxième niveau et ceux de troisième niveau, un syntagme nominal donné amène à plusieurs syntagmes nominaux de niveau supérieur et vice-versa. Ce fait justifie la création, dans la structure de données de la base, des tables spécifiques pour les associations dont les clés composantes sont le code des syntagmes nominaux de niveau inférieur et le code des syntagmes nominaux de niveau supérieur.

6. Conclusion

Bien qu'on ait monté des statistiques descriptives sur les syntagmes nominaux et ses comportements qui ont permis des constatations précises, il est nécessaire de procéder à une évaluation de la maquette et du comportement des syntagmes nominaux en tant que structure d'accès à l'information, avec la participation des utilisateurs. Or, les délais qu'on a eu pour mettre en place ce travail n'a pas rendu possible d'exécution de cette tâche. Cependant, la maquette est prête pour être soumise à l'évaluation proposée. Étant donné la façon dont la maquette a été développée, la construction d'une quelconque autre base de donnés, même dans une autre langue que la langue portugaise est tout à fait possible. Enfin, la maquette peut être un outil d'expérimentation très important pour l'évaluation de cette approche.

« Dans la vie il n'y a pas de solutions. Il y a des forces en marche ; il faut les créer et les solutions suivent. »

SAINT-EXUPÉRY (1900 - 1944)

Conclusion

L'utilisation des syntagmes nominaux en tant que structure d'accès aux informations dans une base de données textuelle se présente comme une alternative aux systèmes traditionnels de banques de données textuelles. Au lieu de ces derniers, où les utilisateurs sont obligés d'apprendre un langage de recherche parfois complexe, la maquette développée guide les utilisateurs, de manière simple et facile, vers les informations qui correspondent à leurs besoins. Ici, il suffit de connaître la manipulation de la souris. En plus, l'interaction que la maquette offre, permet aux utilisateurs de connaître plus rapidement le domaine de la base de données et son contenu.

Au-delà de la facilité d'utilisation de la maquette, ce travail a permis de montrer que : a) les systèmes de recherche d'information utilisant les syntagmes nominaux peuvent être développés au moyen de logiciels de gestion de bases de données commerciaux ; et b) l'arborescence des syntagmes nominaux permet, aux utilisateurs, une navigation plus facile, plus souple et convergente vers l'information.

Le travail n'est pas achevé cependant car on a pu constater au fil du développement de ce mémoire, qu'il y a plusieurs problèmes à résoudre tant au niveau d'extraction automatique des syntagmes nominaux qu'au niveau de la construction de l'interface de recherche d'information.

En ce qui concerne le niveau d'extraction automatique des syntagmes nominaux, on trouve les questions suivantes: a) les calculs des anaphores ; b) les calculs des ellipses ; et c) l'article zéro, principalement dans le cas de la langue portugaise. Par rapport aux

questions 'a' et 'b' la connaissance des plusieurs études en cours sur le domaine ne pourront qu'aider. La question 'c' par contre s'agit d'une particularité courante de la langue portugaise, qui doit certainement être résolue au moment de la mise en place d'un analyseur morpho-syntaxique pour cette langue. Il suffit donc de le développer.

Il faut distinguer encore l'importance de la bonne détermination des centres de syntagmes nominaux et de ses centres complémentaires puisque l'utilisateur commence la recherche à partir de l'ensemble de ces centres. Ainsi, la réussite d'une recherche d'information, utilisant les syntagmes nominaux comme moyen d'accès, est directement proportionnelle à la bonne détermination des centres de syntagmes nominaux.

Quant au niveau de développement de l'interface de recherche d'information, les tâches suivantes sont à mettre en place : a) une étude d'évaluation de cette maquette, avec la participation des utilisateurs, soit avec le corpus existant, soit avec un corpus dans une autre langue ; b) le développement des outils capables de rechercher des syntagmes nominaux de niveau immédiatement supérieur à partir de plusieurs syntagmes d'un niveau donné où il y a une même signification. La tâche 'a' est très importante parce qu'il faut connaître le comportement et l'opinion des utilisateurs sur l'interface de recherche d'information développée. A partir des résultats de cette étude on peut ajuster et améliorer l'interface de recherche d'information en la rendant plus conviviale. Pour résoudre le problème de la tâche 'b', deux solutions se présentent : soit la création d'une structure capable de stocker les synonymes des syntagmes, comme on l'a faite pour les centres des syntagmes, soit l'implémentation de l'opérateur booléen « ou ». L'avantage de la première solution est qu'elle est faite pour le système de recherche d'information et, donc transparente aux utilisateurs. Par contre, la deuxième solution s'avère plus compliquée pour les utilisateurs, étant donné que c'est à eux que revient la tâche d'indiquer quels sont les syntagmes nominaux qui doivent composer l'expression booléenne. Ainsi, la première solution devient la plus convenable.

Bien que l'on puisse créer un système de recherche d'information au moyen des logiciels de gestion de bases de données commerciaux avec une certaine convivialité, il faut

tenir compte des limitations de ces logiciels, soit au niveau du temps de réponse, soit au niveau de la limite de la taille des champs et des opérations de requêtes. Ainsi, l'utilisation de ces logiciels est conditionnée aux caractéristiques de l'application. Pour une application plus professionnel et pour des bases de données de grande taille il faut plutôt développer un système de recherche d'information complet à partir de langages de programmation comme C++, Pascal, etc.

On peut dire finalement qu'il y a déjà des conditions pour la construction d'un système complet de recherche d'information selon l'approche proposée plus haut, c'est-à-dire, un système capable d'extraire les syntagmes nominaux, de construire l'arborescence des syntagmes nominaux, de naviguer dans cette arborescence et de faire la recherche d'information (au moyen d'une interface similaire à celle développée pour ce mémoire). Cela est possible pour deux raisons : 1) il y a, aujourd'hui une offre de plus en plus croissante d'outils de gestion de fichiers, de création d'interfaces avec l'utilisation de fenêtres compatibles avec l'environnement Windows, des langages de programmation chaque fois plus puissants orientés à objet, etc ; 2) du point de vue du traitement de l'information, il y a des travaux développés pour la langue française, dans le cadre du groupe SYDO, et du point de vue de l'interface de recherche d'information, le présent travail vient de compléter la chaîne d'un système de recherche d'information. Ainsi, c'est une question d'ingénierie de logiciel de rejoindre les idées, les travaux développés et de les reprogrammer dans un système intégré de traitement et de recherche d'information.

Ainsi, comme programme pour ma thèse de doctorat, je propose de développer un analyseur morpho-syntaxique pour la langue portugaise et de construire un système de recherche d'information complet, dès l'extraction des syntagmes nominaux jusqu'à l'interface de recherche d'information.

Bibliographie

BOUCHÉ, Richard. « Le Syntagme Nominal, une Nouvelle Approche des Bases de Données Textuelles ». *Meta*. 1989, vol. 34, n°. 3. p. 428-434.

BRITO, Marcílio de. *Réalisation d'un analyseur morpho-syntaxique pour la reconnaissance du syntagme nominal : utilisation des grammaires affixes*. Lyon, 1991. 221 p. Thèse de doctorat d'Etat, Université Claude Bernard, Lyon I.[NC]

CODD, E. F. « A relational model for large shared data banks ». *CACM*. 1970, vol. 13, n° 6.[NC]

CODD, E. F. « Further normalization of the relational model ». *Data Base Systems, Courant computer science symposium 6*, 1971. Rustin R. Editeur, Englewood Cliffs, New Jersey 1972. [NC]

CUNHA, Celso et Lindsey CINTRA. *Nova Gramática do Português Contemporâneo*. Lisboa : Edições João Sá da Costa, 1991. p. 734.

FLUHR, Christian. « Le traitement du langage naturel dans la recherche d'information documentaire ». In.: *Cours INRIA - Interfaces Intelligentes dans l'Information Scientifique et Technique*. 18-22 Mai 1992. p. 103-128.

KURAMOTO, Hélio. *Les Systèmes de Recherche d'Information en Langage Naturel*. Note de Synthèse, Avril, 1995. 50 p.

LAROUK, Omar. *Extraction de connaissances à partir de documents textuels : traitement automatique de la coordination (connecteurs et ponctuation)*. Université Claude Bernard - Lyon I, Thèse de Doctorat, 1994.

LE GUERN, Michel. « Les descripteurs d'un système documentaire : essai de définition », In. : Bès, G.C., Fauchère, P.M., Lagueunière, F. *Actes du Colloque « Traitement automatique des langues naturelles et systèmes documentaires »*. Condenser, supplément I,Université Clermont Ferrand, 1982.[NC]

LE GUERN, Michel. « Un analyseur morpho-syntaxique pour l'indexation automatique », *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1, p. 22-35.

MEKABOUCHE, A. *Un système multi-experts pour la recherche documentaire*. France, 1991. 263 p. Thèse de doctorat d'Etat, Université d'Orleans.[NC]

[NC] Non Consulté.

MEKABOUCHE, A. et BASSANO, Jena-Claude. « Multi-experts Systems for Documentary Research ». *RIAO 91 : Recherche d'Information Assisté par Ordinateur*. Barcelona, 1991. vol. 1, p. 394-413.

METZGER, J-P. *Syntagmes Nominaux et information textuelle : reconnaissance automatique et représentation*. Lyon, 1988. 324 p. Thèse de doctorat d'Etat, Université Claude Bernard, Lyon I.

MICROSOFT CO. *Microsoft Access : Guide de l'utilisateur*. Ireland : 1994. p. 386.

POLLITT, Steven. « CANSEARCH : An expert systems approach to document retrieval ». *Information Processing and Management*. 1987, vol. 23, n° 2. p. 119-138.

SMEATON Alan F. « Prospects for intelligent, language-based information retrieval ». *Online Review*. 1991, vol. 15, n°. 6. p. 373-382.

STRZALKOWSKI, Tomek. « Natural language processing in large-scale text retrieval tasks ». *Text Retrieval Conference (TREC-1)*. Gaithersburg : 1993. p. 173-187.

VETTER, Max. *Modélisation des données : Approches globale et orientée objets*. Paris : Dunod Informatique, 1992.

Bibliographie du Corpus

ARAÚJO, Vânia Maria Rodrigues de. « Informação: instrumento de dominação e de submissão ». *Ciência da Informação.* 1991, vol. 20, n° 1. p. 37-43.

BARRETO, Auta Rojas. « A informação eficaz na empresa ». *Ciência da Informação.* 1991, vol. 20, n° 1. p. 78-81.

CIANCONI, Regina de Barros. « Gerência da informação : mudança nos perfis profissionais ». *Ciência da Informação.* 1991, vol. 20, n° 2. p. 204-208.

DANTAS, Marcos. « Sistemas de Informação : a evolução dos enfoques ». *Ciência da Informação.* 1992, vol. 21, n° 3. p. 192-196.

FERNANDES, Pedro Onofre. « Economia da Informação ». *Ciência da Informação.* 1991, vol. 20, n° 2. p. 165-168.

FURTADO, João Salvador. « Informação técnico-econômica : mais importante do que nunca ». *Ciência da Informação.* 1991, vol. 20, n° 1. p. 20-22.

KLINTOE, Kjeld. « Interação entre empresas com necessidades de informação (=conhecimento) e a estrutura nacional de centros com provisão de conhecimento acumulado : referência especial à estrutura nacional de serviços de informação, documentação e de biblioteca ». *Ciência da Informação.* 1991, vol. 20, n° 1. p. 55-57.

LEITÃO, Dorodame Moura. « A informação como insumo estratégico ». *Ciência da Informação.* 1993, vol. 22, n° 2. p. 118-123.

MAURA, Mariano A. « Consultoria Informatológica em revisão : uma alternativa para serviços de informação personalizados ». *Ciência da Informação.* 1993, vol. 22, n° 3. p. 242-247.

MAURY, Patrick. « Inteligência competitiva e decisão empresarial ». *Ciência da Informação.* 1993, vol. 22, n° 2. p. 138-141.

PAIVA, Denise Werneck de. « Perspectivas do agente da informação no contexto brasileiro ». *Ciência da Informação.* 1990, vol. 19, n° 1. p. 48-52.

PINHEIRO, Marisa Gurjão. « Informação para a Indústria ». *Ciência da Informação.* 1991, vol. 20, n° 1. p. 16-19.

PINTO, Virgínia Bentes. « Informação : a chave para a qualidade total ». *Ciência da Informação.* 1993, vol. 22, n° 2. p. 133-137.

SOUZA, Francisco das Chagas de. « Uso da informação na indústria como paradigma para o desenvolvimento econômico ». *Ciência da Informação.* 1991, vol. 20, nº 1. p. 34-36.

VIEIRA, Anna Soledade. « Conhecimento como recurso estratégico empresarial ». *Ciência da Informação. 1993, vol. 22, nº 2. p. 99-101.*

Matériels utilisés

I - Equipements

a) Micro-ordinateur

Kenitec, microprocesseur Intel PENTIUM, 75 Mhz, 8 Mo de mémoire RAM, 420 Mo de disque dur, 1 lecteur haute densité 3 ½ ", 1 lecteur CD-ROM, 1 souris, écran SVGA couler.

b) Imprimante

Canon, BJC-4000, jet d'encre, couler.

c) *Scanner*

ScanMan Logitech Mod. 256. *Scanner* à main.

II - Logiciels

a) Système d'exploitation

DOS version 6.22 - Microsoft®

Windows version 3.1 - Microsoft®

b) Traitement de textes

Word 6.0a - Microsoft®

c) Gestion de bases de données

Access version 2.0 - Microsoft®

d) Reconnaissance Optique de Caractère (OCR)

OmniPage Direct - Caere®

e) Tableaux et graphiques

EXCEL - Microsoft®

f) Capture d'image Windows

GrabIt Pro - Software Excellence By Design Inc.