AUTOMATISATION DE LA CHAINE DE PRODUCTION TERMINOLOGIQUE

RECONNAISSANCE ET EXTRACTION DES TERMES

S.AIT TALEB* - F. BENJELLOUN**

INTRODUCTION

Dans un monde marqué par le développement de plus en plus fort de la science et de la technique, la communication à travers les frontières linguistiques et, par la même la terminologie, prend une importance croissante dans ce processus global. Le développement des banques de données terminologiques est apparu très tôt comme la seule réponse à la gestion des flots d'informations scientifiques et des milliers de termes qu'elle ne cesse de générer en permettant non seulement d'exploiter et d'organiser l'accès aux données mais aussi d'en maintenir le contenu, de le diffuser et d'en assurer un échange plus efficace et optimal.

Aujourd'hui l'informatique a évolué, vers une plus grande banalisation puisqu'elle s'est introduite dans tous les foyers, qu'elle a envahi tous les domaines de connaissance et qu'elle se rencontre dans tous les environnements de travail. Elle a également évolué vers une plus grande complexification en ce qui concerne le traitement des connaissances, et ce grâce aux acquis des progrès de l'intelligence artificielle qui permettent la constitution de bases de données relationnelles évoluées par le recours à des structures de modélisation et de représentions basées sur les réseaux sémantiques par ailleurs, les techniques hypertextes pour une recherche multicritère et un mode d'interrogation homme-machine qui ce rapproche de plus en plus de la langue naturelle ont également contribué à cette évolution, cette évolution a marqué le travail de la terminologie informatisée. Aussi assistons-nous actuellement à l'émergence d'une nouvelle approche, pour faire face à de nouveaux besoins et à de nouvelles exigences. Deux concepts clefs déterminent aujourd'hui la recherche en matière d'automatisation de la terminologie, le

* Enseignant - chercheur à l'I.E.R.A
** Enseignant - chercheur à l'I.E.R.A

concept des industries de la langue et le concept de terminotique.

Les industries de la langue et la terminologie

Apparu en 1984 dans un article de la revue Brise, le vocable d'industrie de la langue (Ingénierie linguistique) s'est élargi pour signifier «l'ensemble des activités de conception, de production et de commercialisation d'outils et de produits, de services donnant lieu à un traitement automatisé de la langue" [J.C. Corbeil 1990]. Cette nouvelle conception en matière d'informatique et de traitement des langues a quelques peu bouleversé les habitudes des terminolgues. La constitution de banques de terminologie en tant que réservoirs électroniques de données terminologiques, organisées et structurées à des fin de consultation et de diffusion ne représente plus le seul apport de l'informatique à cette discipline. La recherche en terminologie s'oriente vers d'autres outils informatisés pouvant optimiser le travail du terminologue dans tout son processus.

La terminotique

Dans le cadre de ce qui est convenu d'appeler l'industrie de la langue, la terminologie est devenue un concept de recherche particulier de l'informatique linguistique appliquée, un concept programmatique et générateur de produits qui trouvent leur application en terminologie et qui répondent à des besoins d'automatisation des tâches méthodologiques de tout le processus de la chaîne terminologique. Les concepts d'industrie de la langue et de terminotique sont liés, car l'informatique a changé de nature pour traiter des problèmes d'un type nouveau (traitement des corpus textuels, raisonnement incertain sur des données incomplètes, apprentissage automatique...) et est devenu une informatique d'orientation textuelle pour le traitement de l'écrit intervenant ainsi à tous les niveaux du processus terminologique. Dans cet exposé nous allons passer en revue ces différentes étapes et examiner les différents aspects de leur automatisation en s'attardant plus particulièrement sur l'aspect dépouillement, extraction et constitution de la nomenclature. Le schéma ci-après montre les différentes étapes dans la perspective de leur automatisation.

L'AUTOMATISATION DE LA CHAINE DE TRAVAIL EN LEXICOGRAPHIE ET EN TERMINOGRAPHIE

Méthodes traditionnelles	Méthodes informatiques
* Recherche documentaire	
Sélection des documents à dépouiller	1. Téléchargement ou saisie optique ou manuelle pour constituer le corpus écrit. 2. Mise en forme du corpus par un logiciel de traitement de texte.
** Dépouillement	
Choix provisoire de termes	3. Lecture et traitement du fichier texte par un logiciel de découpage de mots/termes.
Choix définitif de termes	4. Recomposition des termes complexes et formes composées
Établissement de la nomenclature	5. Marquage des termes à conserver pour la suite de recherche
*** Contextes	
Cueillette de contextes	6. Lecture et traitement des textes par un logiciel d'indexation, établissement de concordances
Découpage et sélection des contextes	7. Découpage simultané des contextes Délimitation des contextes 8. Importation des contextes dans un S.G.B.D.
Traitement des contextes	9. L'analyse sémantique des contextes à l'aide d'un système de gestion des contextes.

LA CHAINE TERMINOLOGIQUE

Délimitation du domaine et constitution du corpus

Pour un terminologue ou un organisme de mise au point terminologique, le choix du domaine de travail est soit une réponse ponctuelle à un besoin exprimé ou ressenti (les contenus sont alors dépendants de facteurs plus ou moins conjoncturels), soit une recherche programmée qui s'inscrit dans une politique linguistique particulière et prédéfinie pour rattraper le retard accusé dans certains domaines qui restent lacunaires. Dans cette 2ème scénario la terminologie répond à l'hypertechnicité du monde contemporain. Dans un cas comme dans l'autre, le recours à une documentation spécialisée guide le terminologue dans l'exploration du domaine et dans la représentation de sa structure. Le savoir et l'expertise qu'il tire de cette documentation validée par le recours à l'expert lui permettent d'établir **l'arbre de domaine.** Ce dernier est une structuration des différents concepts appartenant à ce domaine, permettant de cerner de façon plus circonscrite les sous domaines ainsi que le volume des notions à traiter. L'étape de la constitution de cette documentation peut être raccourcie par le recours soit à la télématique pour la téléconsultation et le téléchargement à partir de banques de données (bibliographiques, textuelles et terminologiques), soit aux systèmes de lecture optique pour la saisie automatique des textes écrits. Cette saisie automatique de documents imprimés permet au terminologue d'utiliser tous les genres d'écrits pour son travail, d'autant plus, qu'il existe des logiciels commerciaux permettant une bonne fiabilité quant au taux d'erreurs qu'ils génèrent. Pour la langue arabe, les taux d'erreurs restent relativement élevés à cause des spécificités de cette langue: segmentation, ligature, diacritisation, superposition des caractères et formes variables d'un caractère selon sa position dans le mot. Le taux d'erreurs de reconnaissance optique de caractère attribuable à la numérisation influe donc de manière significative sur les résultats obtenus avec le dépouillement automatique ou semi-automatique.

Les textes saisis par lecture optique nécessitent parfois un traitement

préalable (nettoyage, marquage) pour les rendre utilisables par des logiciels dont les formats sont rarement standards ou compatibles entre eux. Par conséquent le recours aux traitements de textes pour accomplir ces tâches de pré-édition est nécessaire.

Reconnaissance des termes et établissement de la nomenclature

Les termes, en tant que signes linguistiques de la dénomination spécialisée désignant un objet concret ou abstrait sans équivoque, ne peuvent différer, quant à leur représentation formelle, des autres chaînes de caractères qui sont les mots. Seule leur appartenance à un domaine spécialisé et leur utilisation par les spécialistes du domaine peuvent être retenues comme critère de différenciation. Néanmoins on peut rappeler quelques spécificités qui caractérisent, du point de vue méthodologique, le terme et qui permettent de le distinguer du mot:

+ Le terme tend à être mono-référentiel dans un texte scientifique et technique appartenant au même domaine et n'admet de synonymie que référentielle.

+ Les termes forment dans chaque vocabulaire scientifique une série, un ensemble dont les éléments sont structurés du fait de leur appartenance à un domaine particulier. Ces éléments sont liés par des relations de types hiérarchique ou associatif qui lient les notions entre elles.

+ Les termes fonctionnent d'une manière particulière sur l'axe paradigmatique. La commutation et la distribution fonctionnent dans une structure calquée du réel.

Constituant l'objet même de la terminologie, la collecte des termes est la première étape dans la chaîne de production de la terminologie. L'opération de dépouillement consiste à relever dans un corpus spécialisé les termes ainsi que les données relatives à ces termes (définition, contextes, termes associés, données nécessaires à l'organisation logique des notions), afin d'établir la nomenclature. Cette étape est cruciale pour le processus terminologique. Elle permet de cerner les notions et de les articuler entre elles à partir d'un examen comparatif des éléments et de proposer une

représentation et une modélisation de la connaissance et du savoir dans le domaine étudié. Elle permet également de déterminer les termes à préviligier, ceux à éviter etc...

L'automatisation de cette étape ne peut être que partielle car l'intervention du terminologue (pour les raisons de caractérisation de termes sus-mentionnées) est indispensable et l'interactivité est requise en raison des exigences de qualité. Par ailleurs, en plus des raisons qui dépendent de la spécificité du travail terminologique, le transfert d'expertise (sous forme d'une terminologie au sens Wusterien d'ensembles bien structurés) pose actuellement plusieurs difficultés :

- Il n'y a pas de théorie générale et unifiée du texte (linguistique, cognitive, documentaire, etc)
- C'est un domaine où le passage de la théorie à l'application entraîne des compromis.
- Désignant des concepts, les termes sont une représentation, ils ne sont pas les "concepts" eux-mêmes dans leur universalité. Ils ne sont que le résultat d'un filtrage particulier (matérialisation linguistique d'un certain nombre de leurs caractéristiques en rapport avec les besoins discursifs) et les critères de reconnaissance par les locuteurs d'une langue spécialisée demeurent mal connus.

Le dépouillement assisté reste un palliatif car il fait gagner du temps au terminologue et permet de traiter un volume textuel plus important et par là même , plus représentatif du domaine.

Gestion terminologique

Il existe actuellement un grand nombre de logiciels dits de terminologie pour la rédaction et la gestion de fiches terminologiques. La plupart de ces produits ont été développés à partir d'outils conçus à d'autres fins et pour des besoins spécifiques (gestion documentaire). Par ailleurs, les logiciels connaissent souvent une utilisation plutôt limitée car leurs principaux utilisateurs se trouvent dans les milieux même de leur conception.

Parmi ces logiciels spécifiques à la terminologie on peut citer, à titre

d'exemple :

AUTOLEX, LEXITERM, NEOLOG, CONCEPTERM, DICOBASE, CAT, KONSULT, LEXICALISE.

AUTOLEX par exemple, est un logiciel de gestion de bases de données terminologiques et en même temps un instrument d'aide à la traduction. Il permet de classer les termes vedettes selon leur degré de fiabilité ou par domaine. Ce système permet également de traiter les sigles et les abréviations.

LEXITERM quant à lui est un logiciel de gestion de bases de données terminologiques multilingues sans limitation du nombre de langues. C'est un système qui permet l'association d'images aux fiches terminologiques.

Quant au **NEOLOG**, c'est un logiciel de confection et de gestion de bases de données terminologiques. Le système accorde une place importante aux champs "constats d'usage" et "notes de recherche", et offre également la possibilité d'établir des liens entre fiches comportant des termes apparentés (synonymes, contraires etc ...).

Le **LEXICALISTE** reste le plus général, il permet de constituer une base ouverte, générique et intelligente : à partir des attributs syntaxiques spécifiés et du réseau sémantique, le système génère une base adéquate à ces spécifications. Il possède un langage de contraintes qui offre la possibilité de laisser le système réduire, à la place de l'utilisateur, la valeur ou les valeurs possibles d'un certain nombre d'attributs en tenant compte des attributs déjà saisis. En plus, des puissants moyens de navigation hypertextuelle dans le dictionnaire qu'il possède, le système comprend plusieurs fonctionnalités, entre autres, un réseau lexical, des liens hypertextuels et la possibilité de navigation sophistiquée dans les réseaux sémantiques et lexicaux.

Exploitation

La plupart des logiciels dits de **terminologie** offrent des stratégies de recherche plus ou moins sophistiquées pour la consultation de la base, et permettent des fonctionnalités, d'impression et d'exportation des données. Quant à l'accès à l'information, les logiciels de terminologie disposent d'un langage d'interrogation à base de termes et d'opérateurs booléens et/ou de

proximité.

Ainsi la recherche peut se faire sur un terme complet ou à partir d'une partie de ses lettres (troncature à droite, à gauche et avec masquage). La recherche peut se faire en texte libre ou sur des champs spécifiés. Des expressions de recherche incluant des opérateurs booléens peuvent être également utilisées (Union, Intersection, Exclusion).

Ces logiciels disposent également de fonctions utilitaires tri selon un ou plusieurs critères, impression et exportation selon différents formats: format prédéfini, format utilisateur ou format ASCII.

Echange

L'intérêt de consultation de plusieurs banques de terminologie existantes n'est plus à démontrer. En effet, les échanges de données entre différents fournisseurs évitent d'avoir à refaire les mêmes travaux. La majorité des logiciels de terminologie offrent des fonctions d'importation et d'exportation de données entre les banques. Néanmoins, l'échange pose quelques difficultés pour la mise à jour des données. La tendance actuelle est à l'établissement de formules de partenariat permettant de mettre en relation les différentes sources d'informations, et au développement d'interfaces pour l'interrogation d'un certain nombre de banques sélectionnées et fusionnées par les liens de la communication. Une solution intermédiaire consiste à utiliser les formats d'échange standards de type SGML, HTML, TEI.

Diffusion

Les banques de terminologie diffusent l'information soit par accès direct aux données soit par le biais de disque optique compact. Différents systèmes de publication ou d'édition électronique sont également utilisés pour augmenter la production et la productivité du secteur de publication, et pour diffuser rapidement les résultats.

EXTRACTION OU ASSISTANCE AU DEPOUILLEMENT DE TEXTES

L'extraction de mots à partir de textes présuppose que ces textes soient déjà disponibles sous format électronique ou qu'ils aient été soumis à une lecture optique qui rend possible leur utilisation

par l'ordinateur. Bien avant de disposer de logiciels d'extraction basés sur l'analyse statistique, ou plus récemment ceux basés sur l'analyse morphologique et syntaxique, le terminologue a cherché, dès l'avènement de logiciel de traitement de textes, à optimiser son travail. Cette approche que l'on peut qualifier de **semi-automatique** concerne l'utilisation des fonctionnalités de traitement de textes et celles des SGBD.

Utilisation de traitement de textes

Dans cette approche, le terminologue, constitue une liste de termes, en délimitant ces derniers à l'aide de symboles conventionnels.

A titre d'exemple nous citons un outil qui a été développé initialement à partir des macros de Word Perfect : IVANOHE[1] Ce logiciel utilise les conventions suivantes

- Les symboles [[ ]] pour délimiter le terme seul et l'importer sans contexte ;
- Les << >> pour délimiter et importer le terme avec son contexte ou la phrase dans laquelle figure le terme marqué ;
- @ marque l'indicatif de page
- Pour les documents bilingues, des indicatifs numériques permettent d'associer un terme et son équivalent s'ils n'apparaissent pas dans le même ordre dans les deux textes

Les versions récentes des logiciels de traitement de textes standard (type WORD, WINTEXT etc ...), offrent la fonctionnalité d'index qui permet au terminologue d'extraire, à partir d'un texte sur support informatique, des termes simples ou composés qu'il aura marqué une seule fois dans son texte.

Cette approche laisse au terminologue le soin d'identifier le terme et lui permet d'accélérer la constitution de la fiche terminologique dont les données doivent être traitées et validés avant d'être versées dans une banque terminologique. La fonctionnalité index est devenue une fonction essentielle dans les logiciels de traitement de textes quelque soit l'environnement de travail .

Néanmoins, l'utilisation de symboles pour délimiter les termes, les contextes et les indicatifs de page présuppose de parcourir tous les textes en les lisant sur

[1] [développé au bureau de traduction au Canada et qui, pour des problèmes de mise à jour de Word Perfect, a été redéveloppé à l'aide d'un langage de programmation indépendant].

écran, et confère à l'opération de marquage un caractère conventionnel qui peut être ressenti par le terminologue comme une contrainte. De plus le défilement sur écran d'un corpus volumineux n'est pas très ergonomique et le versement des résultats de l'index dans un gestionnaire de fiches terminologiques est un processus assez lourd.

Utilisation d'une fonctionnalité de SGBD

Partant de la définition donné par C Gardarin "Un SGBD se compose de trois couches de fonctions :

1. La gestion des récipients de données sur mémoires secondaires.
2. La gestion de données stockées dans les fichiers, le placement et l'assemblage de ces données, la gestion des liens entre données et des structures permettant de les retrouver rapidement.
3. La présentation de données aux programmeurs d'application et aux usagers ayant exprimé leurs besoins en données". [GARDARIN].

Le tri et le classement sont une opération indispensable dans un SGBD. Ils entraînent une réorganisation physique de la base. L'indexation génère un fichier qui ne comporte que les items triés dans la zone choisie préalablement comme clé et chaque item renvoi dans la base, à l'enregistrement qui contient cet item. On peut créer autant d'index que de zones. L'index est indispensable pour l'interrogation d'une base de données.

Certaines données deviennent donc des critères de sélection : toutes les données d'un champ (terme vedette, entrée) prises séparément (inversion) ou considérées comme un bloc. Les SGBD qui traitent des données textuelles ou considérées comme telles permettent d'extraire automatiquement les mots d'un texte, en éliminant les mots vides. Il s'agit dans ce cas d'indexation automatique, car ces mots génèrent un index de base où leur est assigné le nombre d'occurrence, comme l'illustre l'exemple suivant.

EXPAND RADIOACTIF		
REF	INDEX-TERM	TYPE ITEMS
E1	Radio-téléscope	1
E2	Radio-thérapy	2
E3	Radio-transmission	1
E4	Radio-ulna	1
E5	Radio-ulnar	1
E6	Radioactif	32
E7	Radioactifs	3
E8	Radioactinium	1
E9	Radioactive	36
E10	Radioactives	1
E11	Radioactivité	25
E12	Radioactivity	13
E13	Radioaérienne	1
E14	Radioalignement	1
E15	Radioaltimètre	1
E16	Radioautographie	1
E17	Radioautography	1
E18	Radiobalisage	1
E19	Radiobalise	2
E20	Radiocapital	1

EXPAND RADIOACTIF		
REF	INDEX-TERM	TYPE ITEMS
E1	Radideferous	1
E2	Radifère	5
E3	Radii	9
E4	Radio	67
E5	Radio	2
E6	Radio-actif	6
E7	Radio-activation	1
E8	Radio-active	6
E9	Radio-actives	1
E10	Radio-activité	2
E11	Radio-activity	2
E12	Radio-alignement	1
E13	Radio-altimètre	1
E14	Radio-antidiffuseur	1
E15	Radio-astronomie	1
E16	Radio-bicipital	1
E17	Radio-canal	1
E18	Radio-carpien	1
E19	Radio-channel	1
E20	Radio-communication	1

مرقم كيفية		
مرجع	لفظ المشير	عناصر النوع
م1	كَيْفِيَ	17
م2	كَيْفِيَا	1
م3	كَيْفِيَاتَ	1
م4	كَيْفِيَاتُ	4
م5	كَيْفِيَّاتُ	4
م6	كيفية	
م7	كَيْفِيَّةَ	23
م8	كَيْفِيَّةُ	21
م9	كَيْفِيَّةٌ	3
م10	كيكلولية	1
م11	كَيْلَ	2
م12	كَيْلُ	2
م13	كَيْلٌ	9
م14	كَيْلِ	1
م15	كيلْ	1
م16	كَيْلَانِ	1
م17	كَيْلَةَ	1
م18	كَيْلَةٌ	1
م19	كليس	1
م20	كيللو	9

مرقم الكيفية		
مرجع	لفظ المشير	عناصر النوع
م1	الكَيْسِيُ	1
م2	الكَيْسِيُ	7
م3	الكَيْسِيَّةُ	1
م4	الكَيْفَ	3
م5	الكَيْفِيُ	1
م6	الكيفية	2
م7	الكَيْفِيَّةُ	2
م8	الكَيْلَ	2
م9	الكَيْلُ	2
م10	الكَيْلِ	1
م11	الكَيْلُ	1
م12	الكِلْلَتِي	1
م13	الكيلوجرام	7
م14	الكَيْلُوس	1
م15	الكيلوغرام	1
م16	الكيلومتري	1
م17	الكيلومترية	1
م18	الكيلومترية	1
م19	الكيلومترية	1
م20	الكيميائيّ	1

L'indexation automatique dans le cas des SGBD textuels peut être accompagnée d'un ou de plusieurs dictionnaires de base qui fonctionnent comme des dictionnaires d'exclusion, à savoir les dictionnaires de mots vides et les dictionnaires des formes fléchies. Elle peut être également accompagnée d'un thésaurus.

Utilisation des dictionnaires

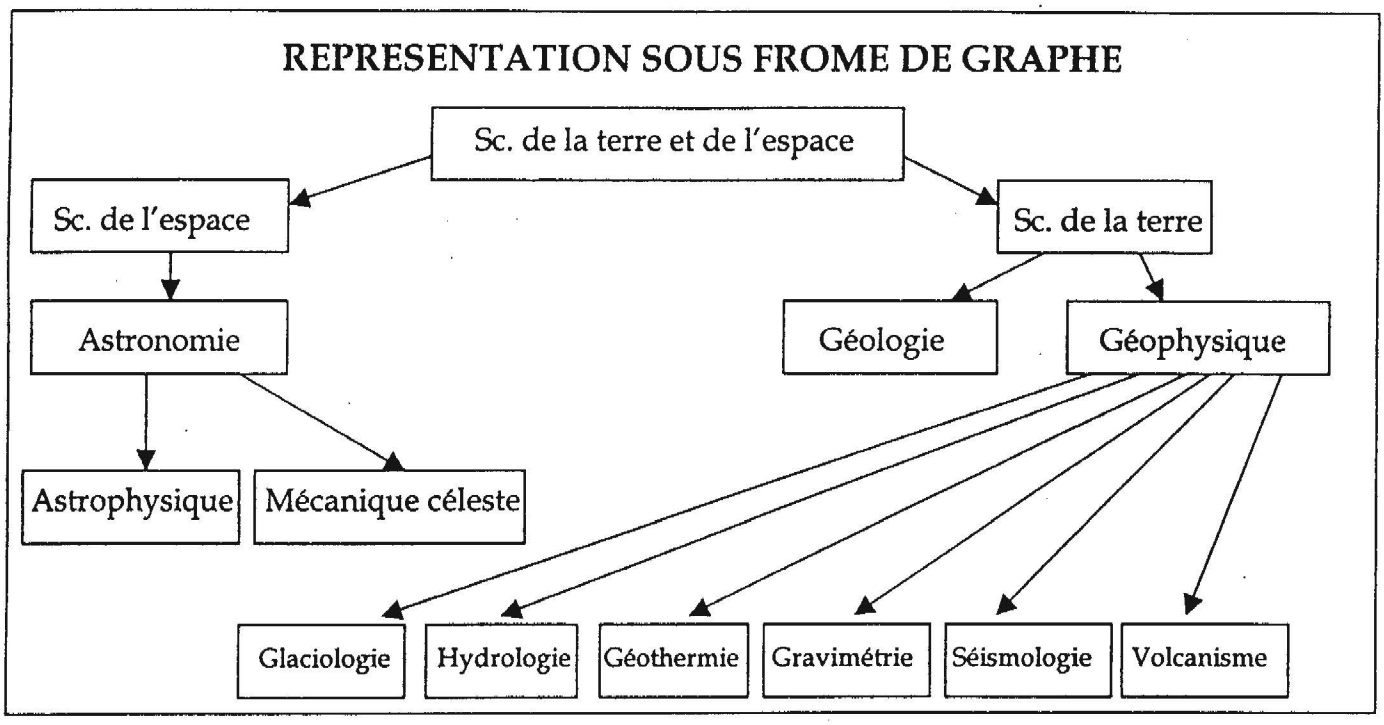
Cette méthode consiste à utiliser un dictionnaire des radicaux et à mettre en place des procédures informatiques efficaces afin de pouvoir traiter n'importe quel mot et le ramener à son radical. Dans cette approche on opère en deux étapes : génération automatique d'un index complet contenant toutes les formes extraites du corpus textuel puis décomposition de la forme pour lui enlever les suffixes et les préfixes.

Utilisation de thésaurus

L'utilisation d'un thésaurus présuppose une procédure particulière qui permet d'étendre la recherche aux termes reliés et d'ajouter aux critères de recherche les termes ayant certaines relations avec les données entrées.

Il s'agit d'un ensemble de termes contrôlés et de concepts liés par différents types de relations stockées dans le thesaurus dont les concepts représentent les nœuds de l'arbre et les termes contrôlés (descripteurs) représentent les feuilles de cet arbre. Chaque concept peut lui même inclure d'autres concepts ou d'autres termes contrôlés. Les relations peuvent être de type générique, spécifique ou de type associatif. Le thésaurus est de ce fait, une structuration d'un domaine du savoir particulier. Le premier niveau de cette structuration décrit les concepts généraux du domaine nommé index. Chaque index est lui même découpé en sous index. Cette structuration peut être modélisée sous forme d'un graphe.

Cette approche **semi-automatique** est donc basée sur la reconnaissance des formes dans une base de donnée textuelle à partir de dictionnaires préalables pour optimiser le rendement de l'extraction (voir schéma ci-après).



Méthode automatique statistique

Avec l'évolution du domaine, des logiciels d'extraction automatique sont apparus et ont évolué du simple segmenteur en passant par des logiciels de statistiques pour aboutir à des logiciels basés sur l'analyse linguistique.

Les segmenteurs

Les segmenteurs découpent le texte en mots à partir d'un texte enregistré sur support informatique. Pour le segmenteur un mot est une suite de caractères comprise entre deux espaces blancs, entre un blanc et un signe de ponctuation ou entre deux signes de ponctuation.

Ce découpage est strictement mécanique et ne tient compte ni des groupes de mots/termes complexes ni de la fréquence des termes.

Ces segmenteurs primaires peuvent être accompagnés de logiciels conçus pour la recomposition des termes composés de manière interactive dans une deuxième étape. Le texte à traiter est défilé dans une fenêtre et l'utilisateur doit indiquer quelles sont les formes complexes qui constituent des termes. Ainsi deux dictionnaires sont constitués, un dictionnaire des termes simples et un dictionnaire des termes composés pouvant être utilisés pour une reconnaissance automatique, cette fois, des termes dans d'autres textes. Voir exemple suivant.

SEGMENTEURS

Mot	Références
le	1
recours	1
à	1
la	1
télématique	2
pour	2
la	2
téléconsultation	3
et	4
le téléchargement	4
à	5
partir	5
de	5
banques	5
données	5
bibliographiques	5
textuelles	6
terminologiques	6

Dictionnaires de mots simples

le
recours
banques
données
textuelles
terminologiques
bibliogiques
bibliographiques
téléconsultation
téléchargement

Dictionnaires de mots compliquées

banques de données
découpages automatiques
banques de données bibliographiques
banques de données textuelles
banques de données terminologiques
données textuelles
données bibliographiques
données terminologiques

Ces segmenteurs posent un certain nombre d'inconvénients à savoir :

- Le défilement des textes sur écran pour la constitution des termes composés est une tâche fastidieuse ;
- la non utilisation de listes d'exclusion des mots sans intérêt tant au niveau des termes simples (mots vides) qu'au niveau des termes composés (locutions adverbiales, prépositives, mots de langue générale) oblige le terminologue à passer en revue toutes formes contenues dans les textes dépouillés.

Les indexeurs et concordanciers

Pour les indexeurs, la récurrence est un des critères essentiels pour déterminer si un mot est représentatif du contenu.

Néanmoins, disposer d'une liste complète des mots d'un texte répertorié et classé par ordre alphabétique ne constitue pas en soi un index, que si cette liste permet de retrouver les différents contextes (chaque mot est accompagné de sa ou ses références du texte). Il convient cependant, de distinguer entre index et concordancier :

Index : Inventaire des mots - formes accompagnées de leur référence.

Concordancier : outil plus sophistiqué, la concordance rassemble toutes les occurrences des mots de texte (le mot est repris autant de fois qu'il figure dans le texte). Chaque occurrence prend place au sein de son contexte original. La constitution d'un concordancier consiste à regrouper toutes les occurrences d'un même mot sous une rubrique, un mot vedette. Ainsi, par rapport à l'index, le concordancier rapproche les contextes dans lesquels s'observe le mot.

Méthode automatique linguistique

Afin de réduire les faiblesses inhérentes aux méthodes statistiques et de récupérer un maximum de termes représentatifs de la terminologie du domaine, des procédures automatiques d'extraction reposant sur des observations linguistiques sont développées. Beaucoup de travaux d'indexation automatique sont fondés sur une extraction des GN. Ce choix se justifie en partie par le fait qu'en langue spécialisée, l'information pertinente est localisée dans les GN et la fréquence de ces derniers en langue spécialisée est importante et élevée.

Cette approche a suscité des études linguistiques et le nombre de logiciels conçus pour analyser et extraire les termes composés est largement supérieur à celui des logiciels traitant les termes simples, malgré le fait que le recours à la sémantique pour leur identification reste difficile (le sens de GN ne dérive pas automatiquement de celui des différents éléments le composant.).

Ces logiciels basés sur l'analyse morphologique et syntaxique constituent actuellement les principaux systèmes d'extraction à des fins terminologiques. Parmi ces logiciels, on peut citer LEXTER, TERMINO, NOMINO, SATO, TOPOCRAF, TERMIS, RTERM.

LEXTER : Logiciel d'extraction de terminologie qui, à partir d'un corpus de textes techniques portant sur un domaine quelconque, effectue une analyse grammaticale de ces textes pour fournir une liste d'unités terminologiques candidates, organisées sous forme d'un réseau hypertextuel. Le terminologue chargé de valider la terminologie proposée par LEXTER peut naviguer au sein du réseau des termes candidats vers les textes et des textes vers les termes candidats qui y'ont été détectés.

NOMINO TERMINO logiciel d'extraction terminologique assistée qui effectue une analyse syntaxique et propose un environnement de dépouillement permettant l'accès aux différentes listes d'unités relevées (Unités complexes nominales, noms, verbes, adjectifs, lexique du texte) ainsi que la consultation de leurs contexte. Le logiciel comporte un gestionnaire de base de fiches avec les fonctions de modification, de suppression, d'interrogation, d'importation et d'exportation de fiches.

La plupart des logiciels passés en revue utilisent des patrons de fouilles basés sur l'observation linguistique, et postulent que certaines formes sont plus terminogènes et productives dans les domaines spécialisés. Ils utilisent des heuristiques linguistiques de type :

- Séquence de la langue qui n'est possible qu'en présence de noms composés D et Prép, Det Qua, Det N Qua , Det V exemple : un à coup, une chasse goupille, un mode à deux temps.
- Probabilité de certaines suites à pouvoir être des noms composés : ellipse du D et N Prép N ex : corrosion par fatigue, juxtaposition de NN exemple : Carte mémoire.
- Etiquetage syntaxique de tous les mots d'un texte français par comparaison avec les dictionnaires électroniques existant, qui identifient les différentes formes fléchies du français et leur apparence syntaxique.
- Repérage de locutions adverbiales
- Indexation par une recherche systémique à l'aide d'une grammaire régulière de formes syntaxique : indexation de figement des noms composés, voir exemple de règles ci-après.

PATRON DE FOUILLE POUR LES TERMES COMPLEXES*

PATRON BASE SUR L'OBSERVATION LINGUISTIQUE

Nde N : gestionnaire de **fichiers**

N Prép N : exploitation en parallèle, calculateur à programme

N Adj : état masqué

erreur logique

N Prép N Adj : calculateur à programme **fixe**

PATRON BASE SUR UNE GRAMMAIRE REGULIERE

Grammaire tirée de :

N1 # Prép # N2 # Prép # N3

1 indique la position dans le GN

≠ désigne une insertion optionnelle (nom, adj, ou mot non reconnu)

Prép # N3 # est optionnel

N1 Adj : appel automatique

N1N2 : carte mémoire

N1 à Det N2 : exploitation à l'alternat

* Applications utilisées dans l'analyse informatique à l'INIST (Centrale documentaire qui produit deux bases de données PASCAL et FRANCIS)

Pour illustrer avec plus de détails cette approche nous présentons le produit TERMINO - NOMINO de dépouillement assisté dont une version démonstration peut être chargée à partir d'internet.

TERMINO-NOMINO est donc un logiciel de dépouillement assisté par ordinateur fonctionnant au départ sur Macintosh. Une version développée pour être opérationnelle sur PC a donné lieu au logiciel NOMINO.

Les textes soumis à TERMINO-NOMINO sont des textes français sur support informatique. Le traitement effectué par TERMINO-NOMINO se compose de trois étapes:

- traitement des marques d'édition, description linguistique et sélection des unités et rédaction des fiches.

Traitement des marques d'édition

Les textes soumis à TERMINO-NOMINO demandent peu de préparation. Les seules règles à respecter sont : un retour chariot ne doit figurer qu'en fin de paragraphe et les traits d'union enfin de ligne doivent être évités.

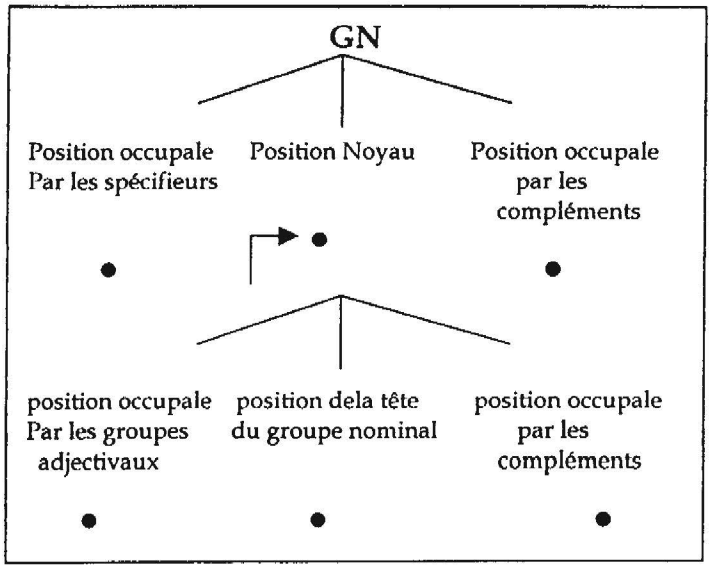
Description linguistique

La description linguistique est effectuée en trois phases

- Phase de catégorisation et de lemmatisation des formes lexicales. Le traitement est effectué forme par forme en indiquant pour chacune d'elle sa catégorie, son lemme et un certain nombre de propriétés (genre, nombre, personne, temps et mode). Ces informations sont nécessaires pour effectuer une analyse syntaxique.

- Phase de l'analyse syntaxique dont le traitement consiste à construire une représentation spatialisée des différentes formes dans l'énoncé. L'analyse s'effectue énoncé par énoncé en s'appuyant sur des propriétés à la fois de structure et des unités lexicales.

- Phase d'extraction des termes composés dépistés autour de certains noms de phrase. Ces termes composés construisent des entités notionnelles ou termes. Voir schéma ci-après.



Les résultats produits par la description linguistique peuvent être filtrés à l'aide de deux procédures fondées sur les opérations de différence et d'intersection appelées, respectivement procédure d'exclusion et procédure d'inclusion. L'exclusion consiste à éliminer des résultats les unités consignées dans un dictionnaire. L'inclusion consiste à ne conserver des résultats que les unités consignées dans un dictionnaire.

La description linguistique produit plusieurs fichiers appelés points de vue correspondant aux noms, adjectifs, verbes, termes composés et au lexique du texte.

Sélection des unités et rédaction des fiches

À partir des résultats de la description linguistique, cette étape consiste à choisir les unités pertinentes, de manière interactive, dans un environnement de bases de fiches. Pour chaque unité, on dispose des différents contextes dans lesquels elle est apparue. Si l'unité est sélectionnée, une fiche est créee automatiquement dans laquelle on inscrit des informations pertinentes (contexte, définition etc ...).

Le processus de sélection d'unités consiste à:

- Choisir un critère fréquentiel à l'aide des opérateurs : tous, égal à, plus petit que, plus grand que, et de la valeur désirée,
- Choisir un point de vue : nom, verbe, adjectif, unité complexe nominale, lexique...
- Choisir une unité dans la liste affichée après le choix du point de vue.

Après le choix de l'unité, le système affiche le contexte de la première occurrence et permet de naviguer dans tous les contextes de toutes les occurrences. Le système permet également d'augmenter le contexte en affichant les phrases précédant et/ou suivant le contexte en cours. Les listes d'unités extraites du texte constituent le point de départ de dépouillement (base de données virtuelle et peuvent permettre de relever des réseaux notionnels partiellement représentés par les familles lexicales que l'on peut explorer).

Exemple : dans le réseau carte, le nœud carte à mémoire conduit lui-même à, carte à mémoire simple. Le nœud mémoire amène aussi à carte à mémoire et supercarte à mémoire, mémoire volalile, mémoire à accès direct, mémoire morte effaçable.

A partir des informations sélectionnées et éventuellement complétées, le terminologue peut enregistrer sa fiche.

Ce produit est présenté à titre d'exemple, d'autres logiciels existent comme : FILCAT, TERMPLUS, TACT.

Ainsi **FILCAT** est un automate qui permet d'extraire les termes complexes, il isole puis filtre des segments complexes de 2 à 9 mots répétitifs d'une base de données textuelles. Les textes sont d'abord indexés par un outil d'indexation de l'analyseur textuel TACT puis un générateur de collocation de TACT produit les segments (2 à 9 mots). FILCAT filtre ce fichier en appliquant des dictionnaires grammaticaux. Voir exemple ci-après.

EXEMPLE DE FICHIER DE COLLOCATIONS GÉNÉRÉ PAR TACT

Reproduction du logiciel

humain	2 création d'un -		2 of -
2 esprit -	2 la création d'un -		2 of system communication
2 l'esprit --	2 de la création d'un -	2 of	system ms communication of
2 de l'esprit -	hyperdocumentation	2 of	systems communication of
hypercard	4 l'-		2 of systems communication of the
2 et -	hyperdocuments		2 of systerns communication of the
hyperdocument	2 – accessibles		acm
5 - électronique	7 d'-		2 eds designing
2 - est	4 des -		2 generation of
2 - et	2 création d'-		2 generation of - systems
2 - 1	2 de la création d'-		2 generation of - systems
10 L'-	hypermedia		communications
2 l' - est	2 - systems		2 generation of - systems
2 l' - et	2 - systems communication		communication of
2 même -	2 - systems communication of		2 generation of - systems
8 un -	2 - systems communication of the		communication of the
2 un - électronique	2 - systems communication of the		2 generation of - systems
6 d'un -	acm		communication of the acm
6 d'un -électronique	2 designing		2 mandl eds designing -
6 de l'--	2 hypertext		2 next generation of --

Pour la langue arabe, il n'existe pas à notre connaissance de produits spécifiquement réalisés pour le dépouillement terminologique. Néanmoins, des analyseurs morphosyntaxiques qui peuvent être intégrés dans l'environnement d'un dépouillement automatique existent au stade prototypique. Nous citons à titre indicatif les approches du D.Y.Hlal. [Applied Arabic Linguistic and Signal Information Processing], pour extraire des concepts (descripteurs en langage documentaire) s'appuient sur une analyse linguistique. Les concepts sont stockés dans un ordre décroissant de pertinence (les critères étant fixés à l'avance : fréquence, charge sémantique posées comme à priori).

Le processus d'analyse morphosyntaxique attache à chaque mot du document à indexer les informations suivantes :

a. Division du mot en éléments morphologiques (éléments qui se combinent pour former le mot).

Préfixe et suffixe CEP + (R - RAT) + ES–

b. Valeur grammaticale (GV) selon la position du mot dans le document. Voir exemple ci-dessous.

وبمدرستهم ← و + ب + (درس – مفعلة) + هم :

عطف – حرف جر – إ سم مضاف – ضمير متصل مضاف إليه

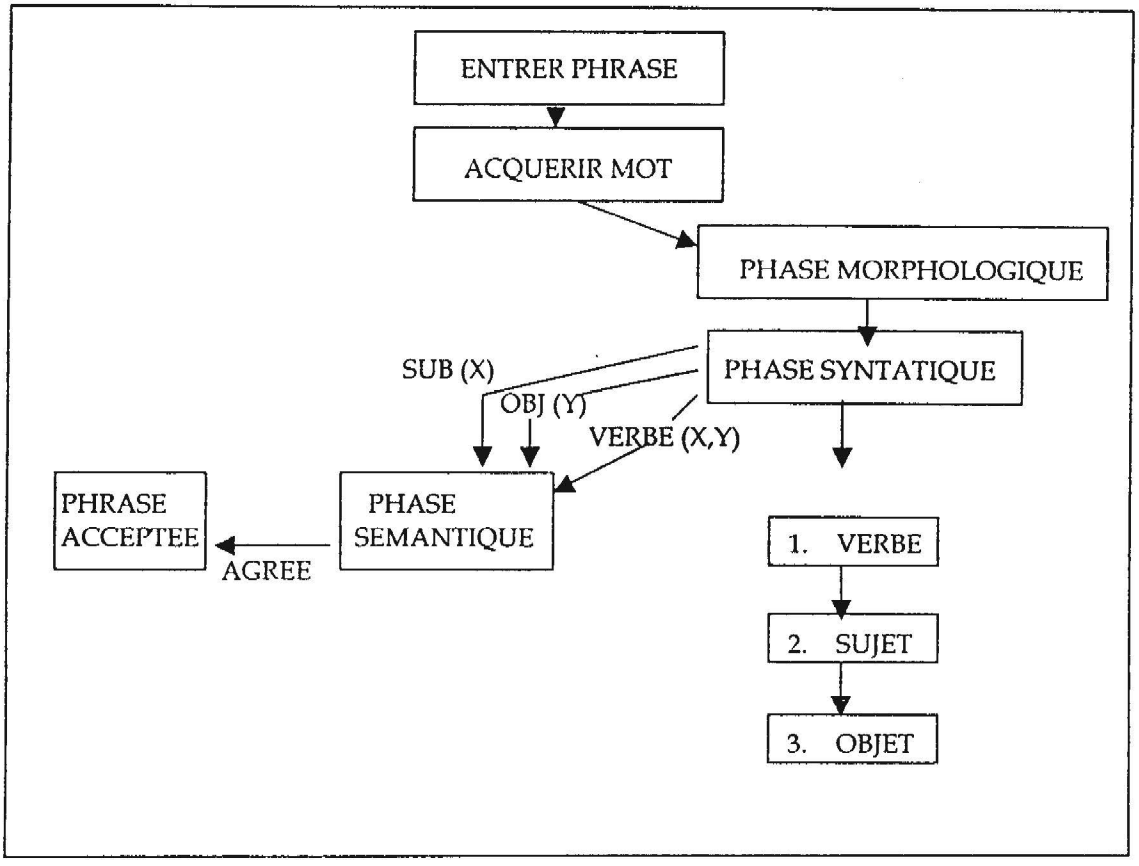
فهم . ← (فهم – فعل) : إسم، فعل ماض، فعل أمر

← ف + (همم – فعل) : عطف – فعل ماض

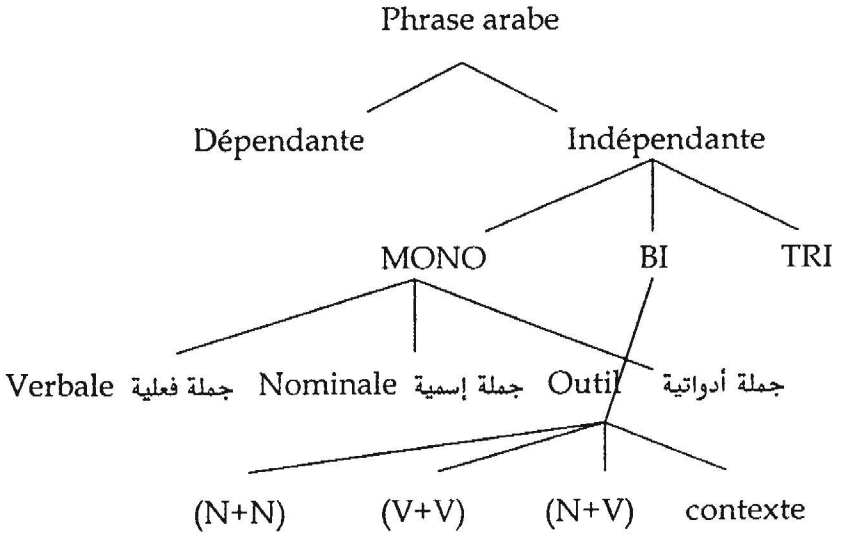
← ف + هم عطف – ضمير منفصل

Nous pouvons signaler également les travaux menés par l'équipe de N.HIGAZI dans le cadre de la constitution d'un système de compréhension de la langue arabe (natural arabic langage understanding system) et notamment la réalisation d'un parseur syntaxique. Ce parseur utilise en plus de l'analyse syntaxique d'une phrase, une base de connaissances sémantiques (prépositions verbales et cas) pour affiner l'inférence sémantique. Ce système est développé en Prolog. Il décompose les phrases en trois types : mono-composées, bi- composées et tri composées. Voir schémas ci-après.

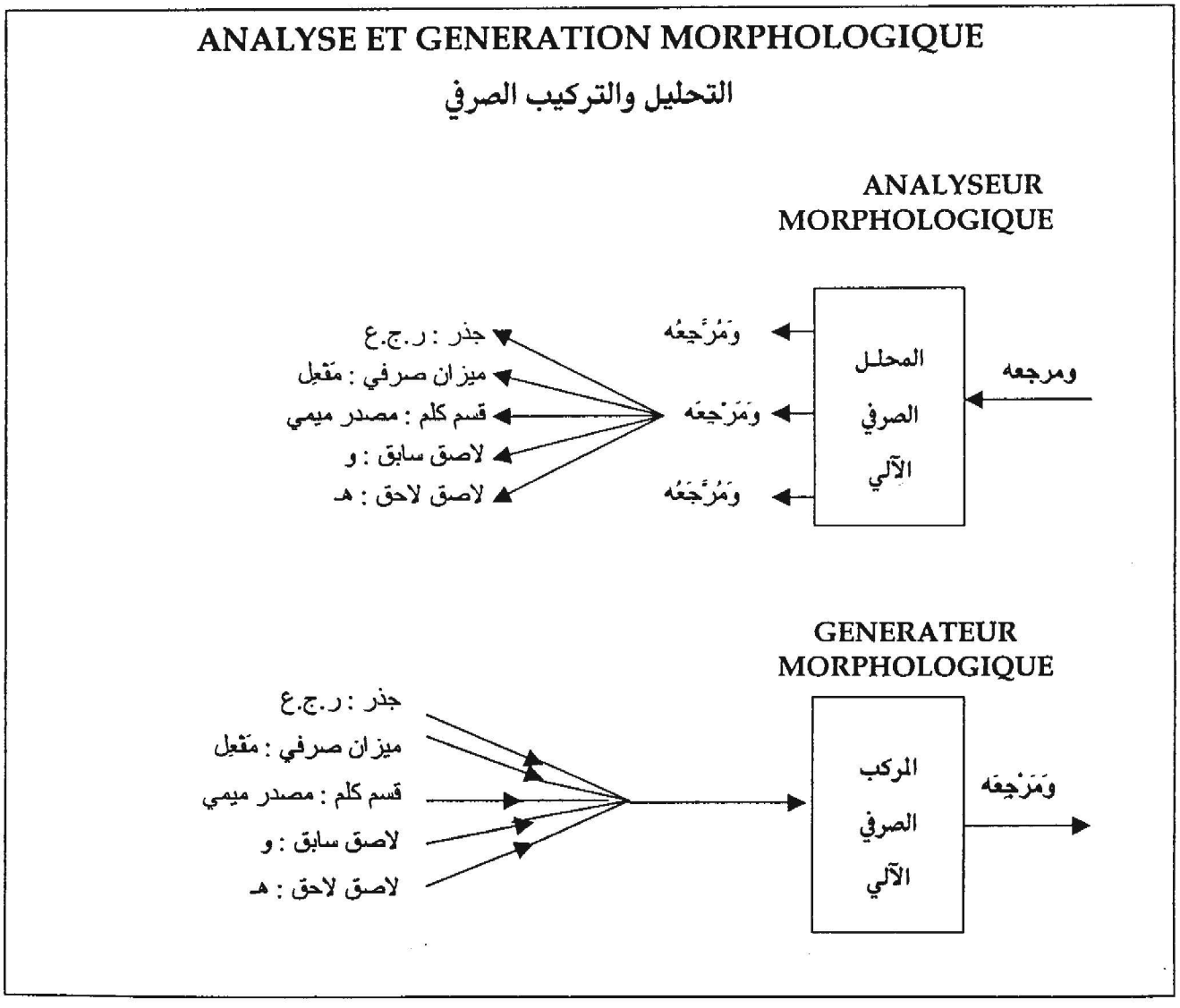
Aperçu sur le système



Classification d'une phrase arabe



Par ailleurs, il convient de signaler les travaux menés par la société SAKHR dans le cadre du système d'extraction des termes arabes (Arabic Text Retrieval System), il s'agit d'un système d'interrogation de type documentaire qui fait appel à un analyseur morphologique lors du traitement des requêtes. Voir exemple ci-après.



CONCLUSION

Les logiciels d'assistance au dépouillement restent inadaptés pour la représentation du sens, car il s'agit ici, non pas d'extraire des chaînes de caractères, mais des termes simples ou composés susceptibles de représenter les concepts du domaine.

Les phénomènes lexicaux morphologiques, syntaxiques, sémantiques et pragmatiques qui sont à l'œuvre dans un texte ne sont pas pris en compte de la même manière. Les descriptions linguistiques proposées, si elles couvrent un niveau (morphologie, syntaxe) occultent souvent les autres niveaux (sémantique, pragamatique).

En effet, pour que l'extraction soit tout à fait automatique, la machine doit être dotée des mêmes connaissances que celles des humains pour garantir une compréhension adéquate des textes. En plus des connaissances linguistiques et l'utilisation de formalisme capable de modéliser un cadre linguistique intéressant, des connaissances supplémentaires sur les domaines couverts (agencement et structuration des concepts, aspects cognitifs, évolution du savoir- etc) ainsi que sur les conditions et les contextes de production de textes (aspect pragmatique et discursif) sont nécessaires.

Ces logiciels se heurtent à des difficultés, de traitement de la polysémie, des formes homographes et de la détermination de la relation inhérente au terme et au concept qu'il dénomme. C'est pourquoi la recherche s'oriente de plus en plus, dans ce domaine, vers la modélisation de systèmes à large couverture linguistique et conceptuelle.

Pour conclure, nous renvoyons à l'évaluation effectuée en 1990 par Salton et Al (Information processing and management Vol. 26), relative à l'efficacité du repérage des différentes méthodes et qui a conclu à l'insuffisance des analyses syntaxiques, à la nécessité de traitement lexico-sémantique et de connaissance qui dépassent le cadre de la phrase.

A la lumière de ce qui vient d'être dit, des questions nouvelles se posent à nous. Une banque de termes c'est aussi un projet de gestion des textes spécialisés et des connaissances, des questions relatives

au choix du système d'un logiciels de gestion des documents électroniques, au choix d'un logiciel d'analyse de texte, au choix d'un dépouilleur de termes, la diffusion du savoir compris dans ces textes sous forme de base de connaissances, sont à l'ordre du jour.