

État de la question des systèmes d'extraction automatique de candidats à terme: vers une proposition intégratrice

Rosa Estopà (rosa.estopa@trad.upf.es)
Jordi Vivaldi (jordi.vivaldi@info.upf.es)
Institut Univesitari de Lingüística Aplicada
Universitat Pompeu Fabra (Barcelona)

À la fin des années quatre-vingts, on ressent la nécessité, depuis des disciplines différentes et avec des finalités différentes, d'extraire automatiquement des unités terminologiques des textes spécialisés; mais ce n'est que dans les années quatre-vingts dix, avec la création de grands corpus textuels informatisés, que les premiers programmes d'extraction sémi-automatique de terminologie commencent à donner des résultats positifs.

Certainement, pendant cette dernière décennie, linguistes computationnels, linguistes appliqués, médiateurs linguistiques (c'est-à-dire, traducteurs, terminologues, interprètes, journalistes scientifiques, etc.) informaticiens, ingénieurs, documentalistes, se sont intéressés, pour des motifs bien différents, à la possibilité de pouvoir isoler, informatiquement, la terminologie à partir de textes.

Les finalités qui mènent ces divers collectifs professionnels à dessiner des outils informatiques qui puissent extraire la terminologie directement des textes sont très diverses:

- la création de glossaires, vocabulaires et dictionnaires terminologiques
- la création de programmes de traduction automatique
- l'indexation de textes
- la création de bases de connaissance
- la création de systèmes hypertextuels
- la création de systèmes experts
- l'analyse linguistique de corpus
- l'enseignement de langues spécialisé, etc.

Qu'est ce qu'un système d'extraction automatique de candidats à terme?

Un système d'extraction automatique de candidats à terme (SEACAT) est un ensemble de programmes informatiques qui **essayent** d'extraire des unités terminologiques à partir d'un corpus textuel informatisé.

J'aimerais souligner le mot **essayer**, parce que dans la réalité tous les systèmes actuels analysent un corpus textuel spécialisé sur un support électronique à partir duquel ils extraient des listes de mots qui sont des **candidats à termes** et que l'utilisateur doit ainsi valider. Pour cette raison, les systèmes de dépouillement de terminologie ne sont pas automatiques mais semi-automatiques, ou autrement dit, il ne s'agit pas de systèmes d'extraction automatique de terminologie, mais de systèmes d'extraction automatique de candidats à terme.

En conséquence, l'objectif de ces systèmes n'est pas de substituer le terminologue, mais d'augmenter la qualité de travail en confiant à l'ordinateur une partie du travail terminologique: le dépouillement terminologique *brut* nécessaire dans tout type d'activité terminologique.

Différents auteurs [Kageura, 1996], [Drouin, 1997], [Estopà, Vivaldi, Cabré, 1998] ont classifié les systèmes d'extraction automatique de candidats à terme en fonction de l'approximation sur laquelle ces systèmes basent la sélection des termes. Ainsi, traditionnellement, on peut distinguer entre:

- les systèmes qui n'utilisent que des méthodes basées sur la connaissance statistique
- les systèmes qui n'utilisent que des méthodes basées sur la connaissance linguistique
- les systèmes qui utilisent des méthodes basées sur la connaissance statistique et aussi la connaissance linguistique.

Les systèmes basés sur la connaissance statistique

En général, les méthodes statistiques assument que les unités lexicales qui concourent avec une fréquence déterminée révèlent un type d'information conceptuelle, en ce sens, on utilise les calculs statistiques pour calculer le degré d'association entre les composants à un candidat à terme. Ces calculs oscillent depuis des simples fréquences jusqu'à des mesures très complexes.

Les méthodes qui se fondent exclusivement sur la connaissance statistique [Salton, 1988], [Evans, 1995], par rapport aux linguistiques, sont conditionnées par la longueur des corpus parce que, si le corpus d'application est petit, il engendra plus de **silence** —nombre de termes qui ne sont pas reconnus du total de termes présents dans un texte—, si le corpus est constitué par des millions d'occurrences il y aura toujours un pourcentage de mots qui, par leur basse fréquence d'usage dans les textes sélectionnés, ne se pourront pas être récupérés. En plus, les méthodes statistiques engendrent aussi beaucoup de **bruit** —nombre de candidats à terme sans la valeur terminologique—, parce que dans les textes spécialisés, mis à part les mots grammaticaux, on trouve autant de mots spécialisés que de mots utilisés avec un sens non spécialisé qui font partie de la langue générale. Parfois, ces mots sans la valeur terminologique apparaissent avec une fréquence d'usage élevée dans les communications spécialisées (*chose, cause, conséquence, hypothèse, sujet, étude, élément, formation, raison, distinction, manière, différence, processus ...*).

Les méthodes statistiques, à la différence des linguistiques, ne permettent pas d'arriver à des généralisations qui contribuent à expliquer des phénomènes du langage naturel, leurs stratégies sont indépendantes du langage. En revanche, entre les méthodes basées sur la connaissance linguistique et les sciences du langage s'établit un circuit de rétroalimentation récursif qui permet d'avancer dans le champ de l'argumentation théorique des phénomènes linguistiques.

Les systèmes basés sur la connaissance linguistique

Les systèmes de base linguistique, comme nous l'avons déjà avancé, utilisent différents types d'information linguistique: il y en a qui se servent de patrons de syntagmes terminologiques, comme le TERMS [Justeson i Katz, 1995]; il y en a aussi qui fonctionnent à partir d'un dictionnaire de mots auxiliaires, d'un dictionnaire terminologique et de règles qui s'appliquent sur des termes reconnus, c'est le cas de FASTR [Jacquemin, 1994]; d'autres utilisent des frontières extérieures du terme, c'est-à-dire, la connaissance linguistique du *non-terme*, comme le LEXTER [Bourigault, 1994], ce type de frontières peut être accompagné d'information typographique sur l'apparition des mots, par exemple DROUIN [Drouin, 1997]; un autre groupe de systèmes se fonde sur des patrons morphostructurels du terme, comme TERMINO [David i Plante, 1991] et d'autres se basent sur une analyse syntaxique détaillée du syntagme nominal, comme le NODALIDA [Arppe, 1995]; etc. Le type de connaissance utilisé fait que les systèmes de cette classe soient, majoritairement, applicables à une seule langue et, ainsi, l'usage dans une langue différente a besoin d'une étude linguistique préalable et, probablement, on doit redessiner grand part du système.

Un des problèmes principaux de tous ces systèmes, qui travaillent à partir d'un type de structure formelle (morphologique, syntaxique et/ou lexicale), est qu'ils engendrent beaucoup de bruit —entre 55% et 75% —, parce que les structures formelles des unités terminologiques ne sont pas exclusives de ce type d'unités linguistiques. Effectivement, les mots proposés par les systèmes comme des unités terminologiques possibles, ne le pas tous sont réellement. Tout au contraire, parfois les mêmes segments peuvent répondre à des unités lexicales avec un usage non spécialisé, ou à des unités phraséologiques spécialisés et encore d'autres sont seulement des segments discursifs.

À la fin du siècle XX, tout le monde partage l'idée que la seule manière de reconnaître et de délimiter, automatiquement, les unités terminologiques d'un texte spécialisé est d'incorporer dans les programmes d'extraction un certain type de **connaissance sémantique**. Dans ce sens et en simplifiant les possibilités réelles, nous disposons de deux grandes approches pour obtenir de l'information sémantique du lexique:

- une approche **fixiste**
- une approche **contextuelle**.

La première approche, de nature fixiste, se fonde sur l'usage des catégories sémantiques d'une source extérieure au corpus textuel de travail. Par exemple, Wordnet¹ ou AlethDic² sont deux systèmes de classification lexicale, de thesaurus lexicals, qui organisent le lexique à partir du signifié des mots et non à partir de leur signifiant. Dans la deuxième approche, de nature contextuelle, il s'agit d'extraire les catégories sémantiques des mots à partir du même corpus de travail.

Dans la première vision, on peut engager l'extracteur de terminologie dessiné par Naulleau (1998):

¹[Beckwith i al. 1990], [Miller, 1990].

²[Naulleau, 1998].

"Contre nos convictions et en raison des problèmes de faisabilité, nous n'adoptons pas le point de vue contextualiste qui aurait pu s'appuyer sur une approche distributionnelle pour définir des catégories sémantiques. Nos catégories sémantiques proviendront donc d'une source extérieure au corpus. Nous avons récupéré les étiquettes sémantiques du lexique d'AléthDic, puis projeté celles-ci sur un nouveau jeu d'étiquettes, plus étroit."

[Naulleau, 1998: 70]

Dans la deuxième optique, on situe le modèle théorique d'interprétation du signifié des séquences proposées par Fabre (1996):

"Notre principale contribution concerne l'élaboration du modèle d'interprétation. Nous avons proposé une mise à plat des règles compositionnelles utilisables pour le calcul sémantique et défini une extension du principe d'attachement d'informations prédictives au nom. Nous avons montré la nécessité de dépasser la limite de ce qui est linguistique pour s'engager dans la prise en compte de données pragmatiques. Nous avons de fait établi des principes d'interprétation pour les séquences sans constituants prédictifs, en nous basant sur certains éléments du lexique génératif de J. Pustejovsky."

[Fabre, 1996:154]

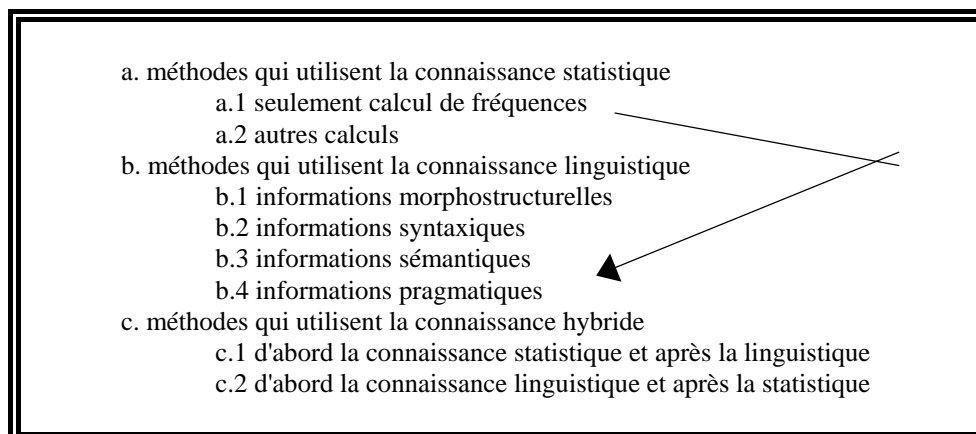
Les systèmes basés sur la connaissance hybride

Finalement, on a les systèmes qui appliquent en même temps la connaissance statistique et la connaissance linguistique. Dans ces systèmes —appelés hybrides—, l'ordre de l'application des types de connaissance est décisif puisqu'il conditionne les résultats. Les méthodes qui appliquent d'abord la connaissance statistique et après la connaissance linguistique présentent les mêmes problèmes de silence que ceux que nous avons déjà commentés par rapport aux systèmes basés exclusivement sur la connaissance statistique, DROUIN [Drouin, 1997]. En revanche, si on utilise la statistique seulement comme un mécanisme complémentaire de la linguistique, à notre avis, les résultats finals peuvent être meilleurs, ACABIT [Daille, 1995] et CLARIT [Evans i Zhai, 1996]. Ainsi, à un moment précis du processus de détection, la statistique peut aider, par exemple, à:

- réaffirmer la condition de terme d'une unité linguistique
- refuser la condition de terme d'une unité linguistique.

À présent, les techniques statistiques fournissent des données par rapport à l'usage des mots et, d'une certaine manière, on peut dire qu'elles suppléent les égards de la compétence pragmatique que tout spécialiste a sur les termes de son domaine.

Si nous tenons compte non seulement du type de connaissance, mais aussi de toute cette diversité d'information utilisée à l'intérieur de chaque type, nous obtiendrons une réponse de classification des systèmes d'extraction automatique de termes plus fine que celle que nous avons proposée auparavant:



Les systèmes d'extraction automatique de *candidats à terme*: *état de la question*

Au moment de dessiner un système d'extraction de terminologie plus complet et plus efficace nous avons fait une analyse à fond des principaux SEACAT actuels pour avoir plus d'éléments de jugement.

Ainsi, nous sommes partis de dix-huit extracteurs différents: ACABIT³, ANA⁴, ATELIER/FX⁵, AUTOLEX⁶, BLANK⁷, CLARIT⁸, DROUIN⁹, FASTR¹⁰, HEID¹¹, LEXTER¹², NAULLEAU¹³, NEURAL¹⁴, NODALIDA-95¹⁵, SBIC¹⁶, TERMIGHT¹⁷, TERMINO¹⁸, TERMS¹⁹, STELLA²⁰.

L'étude des systèmes se centre sur six paramètres:

- les niveaux d'information d'entrée
- les stratégies de reconnaissance de candidats à terme
- les stratégies de filtrage des termes
- les stratégies d'alimentation de connaissance
- l'interaction du système avec l'utilisateur
- les résultats obtenus.

³[Daille, 1994].

⁴[Enguehard et Pantera, 1994].

⁵URL: <http://www.ling.uqam.ca/Ato/FX/AtelierFX.html>

⁶[Planas, 1994].

⁷[Blank, 1995].

⁸[Evans et Zhai, 1996].

⁹[Drouin, 1997].

¹⁰[Jacquemin, 1996].

¹¹[Heid et al., 1996].

¹²Bourigault, 1994].

¹³[Naulleau, 1998].

¹⁴[Frantzi et Ananiadou, 1995].

¹⁵[Arppe, 1995].

¹⁶[Anzaldi, 1996].

¹⁷[Dagan et Church, 1994].

¹⁸[David et Plante, 1991].

¹⁹[Justeson et Katz, 1995].

²⁰[Jacquin et Liscouet, 1996].

Les niveaux d'information d'entrée

Même si SEACAT n'utilise aucun type d'information linguistique, la plupart des extracteurs l'utilise à certains moments du processus. Certains partent d'une liste de mots auxiliaires, d'autres des filtres par catégories grammaticales, mais presque tous utilisent la combinaison d'un analyseur morphologique avec un désambiguateur²¹. Le système dessiné par Naulleau introduit le concept de *syntagme nominal pertinent* pour un usager déterminé, de telle façon qu'à priori on demande à un usager un ensemble de syntagmes nominaux pertinents et un ensemble de syntagmes non pertinents pour ses besoins professionnels et, à partir de ces ensembles, le système extrait des propriétés linguistiques pour construire des filtres négatifs et des filtres positifs.

Le tableau suivant reflète le choix d'information de chaque système:

	Système		Niveau d'information d'entrée			
	nom	listes de mots	analyse morphologique	désambiguateur	corpus d'apprentissage	filtre de catégories
1	ACABIT		X	X		
2	ANA	X				
3	ATELIER/FX		X	X		
4	AUTOLEX	X				
5	BLANK		X	X		
6	CLARIT		X			
7	DROUIN		X	X		
8	FASTR	X	X	X		
9	HEID		X	X		
10	LEXTER		X	X	X	
11	NAULLEAU	X	X	X		
12	NEURAL		X	X		
13	NODALIDA-95		X	X		
14	SBIC	X				
15	TERMIGHT		X	X		
16	TERMINO		X	X		
17	TERMS		X			X
18	STELLA		X			

Les stratégies de reconnaissance de candidats à terme

La reconnaissance et la délimitation des unités terminologiques sont deux des phases les plus complexes de ce type d'application; les programmes analysés s'aident de stratégies différentes pour récupérer les termes, bien qu'aucune des stratégies soit en soi du tout satisfaisante. Ces stratégies sont basiquement:

- les éléments qui jouent le rôle de frontière de mot
- les patrons structurels
- les analyseurs syntaxiques partiels
- les éléments de disposition des mots dans le texte
- les éléments typographiques
- les listes des unités terminologiques
- les profils d'apprentissage

²¹Les auteurs de ces systèmes coïncident à considérer le désambiguateur comme une des sources d'erreur qui fait augmenter l'index de silence.

- les classifications sémantiques du lexique.

Le tableau suivant résume les différentes options adoptées par chaque système analysé:

	Système nom	Délimitation de termes				Désambiguation de structures	
		frontières	patrons	<i>parser</i> ²²	autres	apprentissage	autres méthodes
1	ACABIT		X				-
2	ANA				X		-
3	ATELIER/FX				X		-
4	AUTOLEX	X					-
5	BLANK	X	X				-
6	CLARIT			X			statistique
7	DROUIN	X					-
8	FASTR		X	X	X		-
9	HEID		X				-
10	LEXTER	X				X	
11	NAULLEAU		X	X		X	
12	NEURAL		X				-
13	NODALIDA-95				X		-
14	SBIC	X					manuelle
15	TERMIGHT		X				-
16	TERMINO			X	X		-
17	TERMS		X				-
18	STELLA			X	X		

Les stratégies de filtrage de termes

Avant la dernière phase —explicite ou implicite—, c'est à dire avant de présenter l'ensemble final des candidats à terme des extracteurs de terminologie, a lieu le filtrage de candidats à terme. La représentation suivante concrète le type de filtre que font servir les programmes pour essayer de réduire le bruit initial:

	Système nom	Filtrage de termes					
		manuel	fréquence ²³	linguistique	statistique + linguistique	linguistique + statistique	termes de référence
1	ACABIT					X	
2	ANA						X
3	ATELIER/FX			?			
4	AUTOLEX	X					
5	BLANK		X	X			
6	CLARIT				X	X	
7	DROUIN ²⁴				X		
8	FASTR						X
9	HEID			X			
10	LEXTER			X			
11	NAULLEAU			X			
12	NEURAL					X	
13	NODALIDA-95			X			
14	SBIC	X					
15	TERMIGHT		X	X			
16	TERMINO		X	X			
17	TERMS		X	X			
18	STELLA						X

²²Dans ce contexte, on doit comprendre *parser* comme un outil d'analyse partiel des phrases et jamais comme un outil qui essaie de proportionner une analyse complète et unique de chaque phrase.

²³Nous avons considéré la technique de filtrage de termes avec la fréquence comme un cas particulier, a medio camino entre les méthodes basés sur la connaissance linguistique et les méthodes basés sur la connaissance extralinguistique.

²⁴Ce système incorpore aussi une étape de postprocessament.

Les stratégies d'acquisition

Presqu'aucun des systèmes de terminologie ne profite des résultats obtenus dans l'application du programme, c'est-à-dire que les systèmes n'incorporent pas de techniques de *feedback* et, ainsi, chaque fois qu'on applique de nouveau le programme celui-ci commence à zéro. Seulement deux systèmes, FASTR et ANA, optent pour une stratégie augmentative: à partir d'un ensemble de termes déjà reconnus, le système en reconnaît de nouveaux. Dans ces deux cas, bien que la méthode de reconnaissance soit récursive, les termes identifiés ne sont pas validés avant d'être utilisés dans le cycle suivant et, par conséquent, surgit le problème qu'un segment considéré terminologique de manière incorrecte donne lieu à des termes non valides en cycles postérieurs.

L'interaction du système avec l'utilisateur

Comme nous l'avons dit auparavant, les systèmes d'extraction de terminologie sont semi-automatiques parce qu'à la fin de l'application on arrive à une liste de segments qui doit être validée manuellement par un utilisateur qui possède une compétence cognitive et pragmatique sur le sujet spécialisé. Dans certains cas, les résultats se présentent simplement —par exemple, SBIC—, alors que d'autres systèmes facilitent la tâche de révision à travers soit de la navigation hypertextuelle —LEXTER, ATELIER/FX—, soit de fenêtres avec des multiples contextes pour chaque candidat —NODALIDA, HEID, TERMIGHT, TERMINO—, soit des réseaux sémantiques à partir des termes détectés —ANA, STELLA—, soit en reliant tous les mots d'un texte —ATELIER/FX, FASTR—, ou encore en élaborant un réseau terminologique avec la décomposition des termes en noyaux et en expansion —LEXTER.

Les résultats obtenus

Habituellement, on valorise les résultats obtenus en deux paramètres:

- **le silence**
- **le bruit.**

Dans le premier cas, on valorise le nombre des mots qui, dans le texte, ont une valeur terminologique et que le système n'a pas présentés comme des candidats à terme. Dans le deuxième cas, on mesure le pourcentage d'unités refusées par l'utilisateur du total de candidats à terme présentés par le système, parce que dans le texte elles n'ont pas une valeur terminologique²⁵. Les auteurs des systèmes ne facilitent pas de manière explicite et objective des données sur la réussite de l'application et, dans le cas où nous n'avons pas pu expérimenter avec le système, il est très difficile de connaître le pourcentage de

²⁵Les concepts de *silence* et *recall* et de *bruit* et *precision* donnent la même information depuis des points de vue différents. Ainsi, dans le cas d'un système avec un index de *silence* de 25% il lui correspond un chiffre de *recall* de 75%.

bruit et de silence de chaque système parce qu'il n'y a pas trop de données publiées dans ce sens.

Comme on peut le déduire du tableau suivant, la plupart des systèmes s'appliquent sur des textes dans une seule langue —spécialement en anglais ou en français— et sur des corpus d'un domaine ou d'un sous-domaine très spécialisé:

	Système nom	Corpus d'essai domaine	langue	dimen. [K par.]
1	ACABIT	Télécommunications	français	200 800
2	ANA	Bricolage	français anglais	120 25
3	ATELIER/FX	Medecine	français	?
4	AUTOLEX	?	?	?
5	BLANK	Juridique	allemand	12.000
6	CLARIT	Journalisme	anglais	240 Mbytes
7	DROUIN	Géométrie	français	?
8	FASTR	Medecine	français	1.560
9	HEID	Ingénierie	allemand	35
10	LEXTER	Ingénierie	français	3.250
11	NAULLEAU			
12	NEURAL	Medecine (ophtalmologie)	anglais	55
13	NODALIDA-95	Cosmologie Ingénierie Ind. de l'automobile Journalisme	anglais	20
14	SBIC	Environnement	italien	?
15	TERMIGHT	Informatique	anglais	?
16	TERMINO	Medecine	français	?
17	TERMS	Métallurgie Ing. spacial Ing. nuclear statistique Sémantique Cromatographie	anglais	? ? ? 2,3 6,3 14,9
18	STELLA	Documents d'Internet	anglais français	?

Les systèmes d'extraction automatique de candidats à terme: conclusions

Les paramètres analysés dans le paragraphe antérieur ont mis en évidence les éléments substantiels qui définissent les extracteurs de terminologie et leurs limitations; pour conclure, on peut résumer ces caractéristiques, aux points suivants:

a) **L'efficacité** des systèmes d'extraction présentés est très subjective; dans la plupart des cas, l'auteur n'explique pas les résultats finals d'une manière claire et quantifiable. A l'heure d'évaluer les résultats on doit aussi penser que ces systèmes ont été mis à l'épreuve avec des corpus très **petits** —de 2,3 a 12 quilomots— et hautement **spécialisés** —qu'il s'agisse du sujet ou du niveau de spécialisation. Ce manque de données rend difficile l'évaluation et la comparaison de ces systèmes, même si ceci n'empêche pas de considérer comme intéressantes les solutions proposées par quelques points déterminés.

b) Aucun des systèmes d'extraction d'unités terminologiques est totalement **satisfaisant**. Cette affirmation repose basiquement sur deux faits: d'un côté, tous les systèmes produisent une quantité trop grande de **silence**, surtout ceux de base

statistique; d'un autre côté, tous engendrent une quantité très élevée de **bruit**, surtout ceux de base linguistique qui utilisent une série de **patrons** morphosyntaxiques pour identifier les termes complexes parce qu'ils se fondent seulement sur l'aspect formel de l'unité terminologique.

c) Vu le bruit généré, tous les systèmes d'extraction proposent des **listes de candidats à terme** qu'on doit accepter ou refuser manuellement à la fin du processus. Par conséquent, nous pouvons affirmer que tous les systèmes informatiques d'extraction de terminologie actuels sont **semi-automatiques**.

d) La plupart des systèmes s'appliquent à **une seule langue**, habituellement le français ou l'anglais. Il n'y a aucun système dessiné pour le catalan ou pour le castillan.

e) Tous ces systèmes évalués se centrent exclusivement sur le **syntagme nominal**, aucun système d'extraction ne fait allusion aux **syntagmes verbaux**. Ce fait est motivé par le pourcentage élevé de syntagmes nominaux terminologiques qui s'utilisent dans les textes spécialisés. Mais, on ne peut pas oublier que dans tous les domaines spécialisés il y a aussi des **termes simples** et des combinaisons spécifiques de base verbale, bien que leur pourcentage soit inférieur.

La reconnaissance automatique des termes simples présente divers problèmes: d'abord, leur structure ne présente aucune spécificité —à l'exception de quelques affixes dans certains domaines, par exemple *-itis* en médecine, *-ème* en linguistique, *-asa* en biochimie—; sémantiquement, ce sont des unités beaucoup plus polysémiques que les unités complexes —nous pensons à des mots comme *poid*, *fenêtre*, *clé*, *base* ou *élément*— et, en plus, la fréquence n'est pas toujours un indicateur du caractère terminologique d'un segment. Pour reconnaître automatiquement les termes simples —et aussi les complexes— il est indispensable d'avoir des études sur la sémantique lexicale. En conséquence, actuellement, les systèmes d'extraction de terminologie se concentrent exclusivement sur les unités complexes nominales.

f) Aucun système d'extraction de terminologie ne fait allusion, donc, à la délimitation entre les collocations nominales et les unités terminologiques nominales syntaxiques: pas plus qu'à l'extraction de la phraséologie verbale.

g) Seul un des systèmes analysés utilise l'information sémantique pour reconnaître et délimiter les unités terminologiques bien qu'il y ait certains systèmes qui font servir des heuristiques avec des filtres lexicosémantiques. Il faut dire que les systèmes de représentation de la sémantique lexicale sont actuellement bien peu.

h) Aucun système n'utilise à fond les caractéristiques combinatoires propres des termes des langues du spécialité attachées à un sujet. Il serait important de disposer de plus d'études sur des types de restrictions que présentent les unités terminologiques en relation:

- au champ conceptuel
 - au type de texte.
- i) Quelques stratégies utilisées par différents systèmes semblent particulièrement intéressantes:

- l'usage des règles heuristiques, tant pour l'unité terminologique comme pour celle qui ne peut jamais l'être
- l'élaboration de réseaux d'expansions et de noyaux de termes complexes
- la réutilisation de termes reconnus
- l'analyse partielle des phrases pour obtenir SN potentiellement terminologiques
- l'extraction de relations sémantiques entre les termes —ou leurs constituants
- l'importance des caractéristiques de disposition des unités terminologiques dans les textes, etc.
- la combinaison de plus d'une stratégie
- l'usage d'un dictionnaire avec de l'information sémantique sur le lexique.

Pour améliorer ces systèmes d'extraction de terminologie et obtenir une réduction du silence et du bruit, il faudrait approfondir surtout deux types d'études. D'un côté, il faudrait plus **d'études linguistiques** sur:

- les relations sémantiques des termes
- les relations sémantiques entre les différents constituants d'une unité terminologique
- l'interprétation sémantique du lexique à partir de corpus textuels
- la représentation sémanticolexique
- les restrictions des unités terminologiques dans un domaine spécialisé concret et dans un type de texte concret
- l'étude de toutes les catégories grammaticales susceptibles d'être des termes dans les différents domaines spécialisés
- l'influence de la fonction syntaxique des syntagmes terminologiques dans les textes
- la sémantique de contexte des unités terminologiques
- les relations entre les termes et leur disposition dans les textes
- les relations entre langues différentes des termes d'un même réseau conceptuel.

Et, d'un autre côté, il faudrait travailler sur l'idée de **systèmes informatiques** qui:

- alternent de manière plus active les méthodes statistiques avec les linguistiques
- améliorent les mesures statistiques
- combinent plus d'une stratégie
- améliorent les interfaces pour favoriser l'interaction machine/usager.

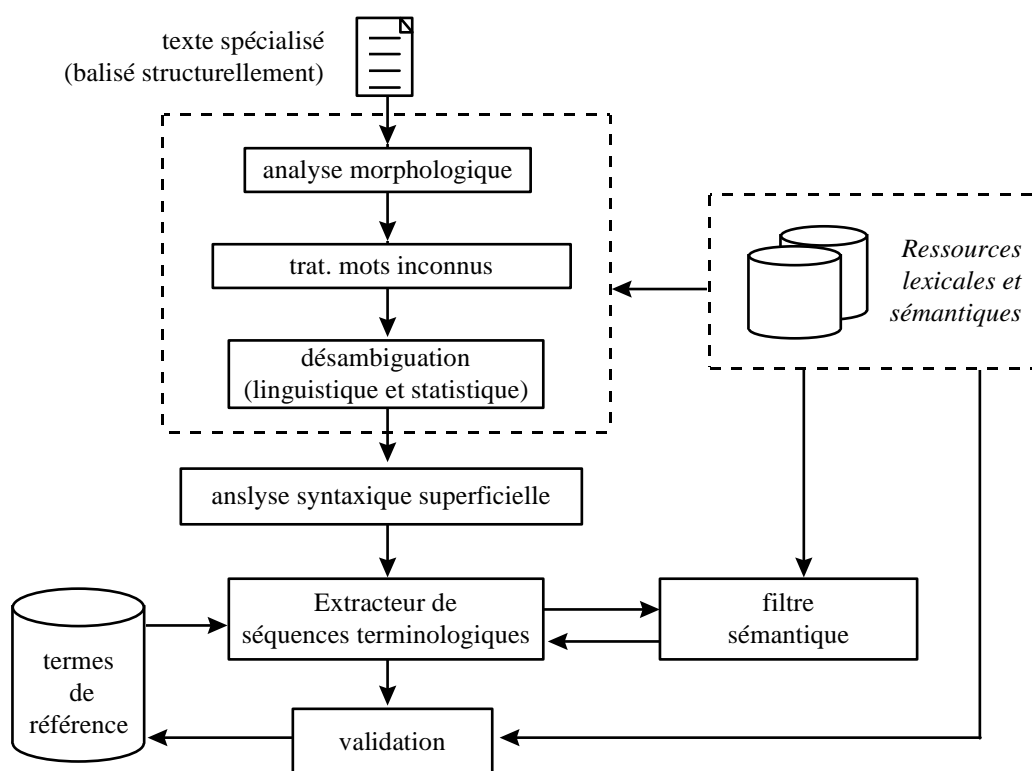
En définitive, si on veut avancer dans le champ de l'extraction automatique de terminologie, les méthodes linguistiques et certaines techniques statistiques doivent interagir activement. Elles ne sont pas, ainsi, approches excluantes, mais complémentaires: la statistique peut compléter la linguistique parce qu'elle aide à connaître l'usage réel qu'on fait des mots. L'objectif final de ces améliorations serait de réduire au maximum le silence et le bruit, de manière à ce que le processus de dépouillement terminologique à partir de corpus textuels spécialisés arrive à être le plus automatique et le précis possible.

Dans la cadre de l'IULA, parallèlement à des recherches linguistiques sur des textes spécialisés pour trouver de régularités sémantiques, morphologiques, syntaxiques, typographiques, etc., nous sommes en train d'élaborer un SEACAT, qui sera le premier pour le catalan et pour le castillan, qui intégrera des stratégies différentes avec le recyclage des outils existants.

Le système que nous proposons essaie de donner une solution à certains problèmes que nous avons exposés antérieurement. L'idée est d'utiliser un système linguistique basé sur de patrons morphosyntaxiques, mais qui réduise le bruit propre de ce type de systèmes. Ainsi, la proposition consiste à:

- intégrer diverses stratégies,
- approfondir l'étude des caractéristiques spécifiques des termes
- rapprocher les candidats à terme à la notion de terme.

Le schéma suivant montre la première maquette de notre système:



Pour y parvenir, nous partirons des textes balisés structurellement avec le SGML. Ces textes seront analysés et désambiguïsés morphologiquement avec des processus propres de ces systèmes même si nous chercherons à préciser le traitement des mots inconnus et à améliorer le processus de désambiguation dans un domaine concret. A continuation, nous ferons une analyse syntaxique superficielle qui nous permettra de connaître les structures sous-jacentes de chaque phrase du texte.

Un des objectifs de cette proposition est la combinaison de stratégies; en ce sens, nous pensons que la réutilisation de termes avec un filtre sémantique peut être un bon chemin pour obtenir un rendement global positif du système. Ce filtre fonctionnera à partir des propositions de l'extracteur de séquences et de la information extraite des bases de données sémantiques du domaine traité. De cette manière, nous pensons réduire sensiblement le bruit et présenter à l'usager une liste plus *propre*.

En resumé, le but de cet exposé a été de montrer un panorama général du champ de l'extraction automatique de termes. Nous avons vu qu'il y a beaucoup de SEACAT surtout pour le français et l'anglais. Nous avons analysés de façon détaillé les principaux SEACAT pour obtenir un ensemble de critères de jugement pour la conception d'un logiciel d'aide au dépouillement terminologique. Nous pensons que les SEACAT sont d'une grande aide pour tout type de travail terminologique, mais ils ne deviendront efficaces qu'avec la compénétration et la coopération actives de toutes les disciplines en relation avec la terminologie.

Bibliographie de référence

1. ARPPE, A (1995): "Term extraction from unrestricted text". Lingsoft Web site. (<http://www.lingsoft.com>).
2. AHMAD K. i al. (1996). "Engineering Terminology - A case for a linguistically-informed terminology database". *TKE '96: Terminology and Knowledge Engineering*. Berlín: Indeks Verlag. Pag. 166-178.
3. BACH, C. i d'altres (1997). *El Corpus de l'IULA: descripció*. Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada . Papers de l'IULA, Informes, 17.
4. BLANK, I. (1995) "Méthodes pour l'extraction de terminologie bilingue". *Actes de les IVèmes Journées scientifiques du Réseau Lexicologie, Terminologie, traduction*. Lió. [À éditer]
5. BORDONI, L.; ANZALDI, C. (1996). *Prototipo di thesaurus per l'energia e l'ambiente tramite il sistema SBIC*. Informe RT, STUDI, 1996, 1.
6. BOURIGAULT, D. (1993). "Analyse syntaxique locale pour le reperege de termes complexes dans un texte". *TAL*, 2. Pag. 105-117.
7. ---. (1995). "Conception et exploitation d'un logiciel d'extraction de termes: problèmes théoriques et méthodologiques". *Actes des IVèmes Journées scientifiques du Réseau Lexicologie, Terminologie, Traduction*. Lió, [À éditer].
8. ---. (1994). *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes*. Paris, École des Hautes Études en Sciences Sociales. [Thèse doctorale].

9. ---. (1996). "LEXTER, a Nature Language Processing Tool for Terminology Extraction". Actes du 7th EURALEX International Congress. Göteborg, [À éditer].
10. BOURIGAULT, D.; CONDAMINES, A. (1995). "Réflexion sur le concept de Base de Connaissances Terminologiques". Actes des 5^{èmes} Journées Nationales du PRC GDR Intelligence Artificielle [À éditer].
11. CABRÉ, M. T. (1992): *La terminologia. La teoria, els mètodes, les aplicacions*. Barcelona. Empúries.
12. CHURCH, K. W. (1989). "Word association norms, mutual information and lexicography". Actes du 27th annual meeting of the ACL. Vancouver. Pag. 76-83.
13. CONDAMINES, A. (1995). "Terminology: new needs, new perspectives". *Terminology*, 2, 2. Pag. 219-238.
14. DAGAN I.; CHURCH K. (1994). "Termight: Identifying and translating technical terminology". Actes de la Fourth Conference on Applied Natural Language Processing. Stuttgart.
15. DAILLE, B. (1994). *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Université Paris VII. [Thèse doctorale].
16. DAILLE, B. i al. (1996). "Empirical observation of term variations and principles of their description". *Terminology*, 3, 2 p.197-257.
17. ---. (1995). "Repérage et extraction de terminologie par une approche mixte statistique et linguistique". *TAL*, 36,1-2. Pag. 101-118.
18. DAVID, S. (1995). *Les Unités nominales polylexicales. Éléments de description et reconnaissance automatique*. Université Denis Diderot. Paris VII. [Thèse doctorale].
19. DAVID, S.; PLANTE. Pag. (1991). "Le progiciel TERMINO: de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes". Actes du Colloque de Montréal *Les industries de la langue: perspectives des années 1990*, 1991, 1. Pag. 71-88.
20. DROUIN, P. (1996). "Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme". [À éditer]
21. ESTOPÀ, R. (1996). *Les unitats terminològiques polilexemàtiques en els lèxics especialitzats (dret i medicina)*. Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada. [Thèse de DEA]
22. ESTOPÀ, R.; VIVALDI, J.; CABRÉ, M. T. (1998). "Sistemes d'extracció automàtica de candidats a terme. Estat de la qüestió". Papers de l'IULA, Série Informes, 22. Pàgs. 1-68.
23. ENGUEHARD, C.; PANTERA, L. (1994). "Automatic Natural Acquisition of a Terminology". *Journal of Quantitative Linguistics*, 2, 1. Pag. 27-32.
24. EVANS, D. A.; ZHAI, C. (1996). "Noun-phrase Analysis in Unrestricted Text for information retrieval". Actes del 34th Annual Meeting of ACL. Santa Cruz: University of California, 1996. Pag. 17-24.
25. FRANTZI, K.; ANANIADOU, S. (1995). "Statistical measures for terminological extraction". Working Papers du Department of Computing of Manchester Metropolitan University.
26. GUILLET, A. (1990). "Reconnaissance des formes verbales avec un dictionnaire minimal". *Langue française*, 87. Pag. 52-58.

27. HABERT, B.; NAULLEAU, E.; NAZARENKO, A. (1996). "Symbolic word clustering for medium-size corpora". Actes de *Coling'96*. Pag. 490-495.
28. HEID, U. i al. (1996). "Term extraction with standard tools for corpus exploration. Experience from German". *TKE '96: Terminology and Knowledge Engineering*. Berlín: Indeks Verlag. Pag. 139-150.
29. HULL, D. i al. (1996). "Xerox TREC-5 site report: routing, filtering, NLP and Spanish tracks". Actes du *TREC-5*.
30. JACQUEMIN, C. (1994). "Recycling Terms into a Partial Parser". Actes de la 4th *Conference on Applied Natural Language (ANLP'94)*. Stuttgart. Pag. 113-118.
31. JACQUEMIN, C. (1996). "What is the tree we see through the window: A linguistic approach to windowing and term variation". *Information Processing & Management*, 32, 4. Pag. 445-458
32. JACQUIN, C.; LISCOUET, M. (1996). "Terminology extraction from texts corpora: application to document keeping via Internet". *TKE '96: Terminology and Knowledge Engineering*. Berlín: Indeks Verlag. Pag. 74-83.
33. JUSTESON, J.; KATZ, S. (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text". *Natural Language Engineering*, 1, 1. Pag. 9-27.
34. KAGEURA, K.; UMINO, B. (1996). "Methods of Automatic Term Recognition". Working Papers: *National Center for Science Information Systems*. Pag. 1-22.
35. KARLSSON, F.(1990). "Constraint grammar as a framework for parsing running text". Actes de la 13th *International conference on computational linguistic*, vol. 3. Pag. 168-173.
36. LAURISTON, A. (1994). "Automatic recognition of complex terms: Problems and the TERMINO solution". *Terminology*, 1, 1. Pag. 147-170.
37. L'HOMME, M. (1996). "Sélection des prépositions dans les termes complexes Nom (Prép.) Nom à partir de leur structure conceptuelle". *Cahiers de Lexicologie*, 68, 1. Pag. 25-43.
38. MILLER G. i al. (1993). *Introduction to WordNet: An On-line Lexical Database*. University of Princeton. [Working Paper]
39. MORIN, E. (1995). *Acquisition automatique de liens sémantique dans les corpus de textes: Application à l'hyponymie*. Université de Nantes. [Mémoire de DEA d'informatique]
40. NAULLEAU, E. (1998) *Apprentissage et filtrage syntaxico-sémantique de syntagmes nominaux pour la recherche documentaire*. Université Paris VIII [Thèse doctorale].
41. OTMAN, G. (1991). "Des ambitions et des performances d'un système de dépouillement terminologique assisté par ordinateur". *La banque des mots*, 4. Pag. 59-96.
42. PERRON, J. (1991). "Présentation du progiciel de dépouillement terminologique assisté par ordinateur: Termino". Actes du Colloque de Montréal *Industries de la langue: perspectives des années, 1990*. 1991, 2. Pag. 715-755.
43. PINEIRA-TRESMONTANT, C. (1992). "Reconnaissance automatique des unités syntagmatiques". [Working Paper] Pag. 1-13.
44. PLANAS, A. (1994). "AUTOLEX: Sistema para la gestión de bases de datos terminológicas y herramienta para la traducción asistida por computadora". *Ciencias de la información*, 25.
45. SHIEBER, S. N. (1986). "An Introduction to Unification-Based Approaches to grammar". *CSLI Lecture Notes*, vol 4, Chicago University Press.

46. SMADJA, F. (1991). *Extracting collocations from text. An application : language generation*. Columbia University. Department of Computer Science. [Thèse doctorale]
47. VOUTILAINEN, A. (1993). "NPtool, a detector of english noun phrases". Actes del *Workshop on Very Large Corpora*. Columbus: Ohio State University.
48. ZHAI, C. i al. (1996). "Evaluation of syntactic prase indexing - CLARIT NLP track report". Actes du *TREC-5*.