Sistemes d'extracció automàtica de (candidats a) termes: Estat de la qüestió

R. Estopà, J. Vivaldi, M.T. Cabré

Sèrie Informes, 22

Barcelona
Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada
1998

Direcció de les Publicacions de l'IULA: M. Teresa Cabré

Primera edició: 1998

La Rambla, 30-32
08002 Barcelona

Dipòsit legal: B-34.230-2002

Sistemes d'extracció automàtica de (candidats a) termes:

Estat de la qüestió[1]

Rosa Estopà
rosa.estopa@trad.upf.es

Jordi Vivaldi
jorge.vivaldi@info.upf.es

M. Teresa Cabré
teresa.cabre@trad.upf.es

Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona

Des de l'aparició de TERMINO l'any 1990 fins avui dia s'han portat a terme una sèrie de projectes per dissenyar diferents tipus d'extractors automàtics de terminologia, però malgrat la gran quantitat d'estudis que s'estan realitzant en aquesta línia, els resultats no són del tot satisfactoris.

Aquest article presenta una anàlisi dels principals sistemes d'extracció automàtica de terminologia amb la finalitat de dibuixar un panorama general de l'estat de la qüestió. L'estudi comença amb la descripció de diversos sistemes d'extracció de terminologia; en el segon apartat, es comparen els diferents sistemes. L'informe finalitza amb unes conclusions sobre els sistemes analitzats i sobre els criteris que podrien guiar una proposta d'un **sistema integrador** d'extracció de terminologia.

Since the presentation of TERMINO in 1990 until today there have been a number of projects aiming to design several types of terminology extractors. However, despite the great deal of studies that are being carried out in this way, the results are not satisfactories.

This article analyses the main systems of terminology automatic extraction so as to give an overview of the current state of the art. It starts with the description of some systems of terminology extraction and follows with their comparison. Finally we come to some conclusions about the systems that are analysed and the criteria that could lead to the proposal of a methodology-combined system of terminology extraction.

[1] Aquest article s'ha realitzat en el marc del projecte de recerca *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica* (PB-96-0293).

Índex

1. INTRODUCCIÓ 1

2. DESCRIPCIÓ DELS PRINCIPALS SISTEMES D'EXTRACCIÓ DE TERMINOLOGIA . 7

2.1 ANA 7
2.2 Atelier FX 9
2.3 Autolex 11
2.4 Blank 12
2.5 CLARIT 14
2.6 Daille-94 17
2.7 Drouin 19
2.8 FASTR 21
2.9 Heid 25
2.10 LEXTER 27
2.11 NEURAL 31
2.12 NODALIDA-95 33
2.13 SBIC 35
2.14 Termight 36
2.15 TERMINO 38
2.16 TERMS 42
2.17 STELLA 44

3. ESTUDI COMPARATIU 46

3.1 Nivells d'informació d'entrada que utilitzen 47
3.2 Estratègies de delimitació de termes 48
3.3 Estratègies de filtratge de termes 49
3.4 Estratègies d'adquisició 50
3.5 Classificació del termes reconeguts 51
3.6 Interacció amb l'usuari 51
3.7 Resultats obtinguts 52
3.8 Productes comercials 53

4. CONCLUSIONS 54

4.1 En relació als sistemes analitzats 54
4.2 En relació a una proposta d'un sistema integrador de detecció automàtica de terminologia 56

5. BIBLIOGRAFIA 62

1. Introducció[2]

A finals dels anys vuitanta es fa evident la necessitat, des de disciplines diferents i amb finalitats diferents, d'extreure automàticament les unitats terminològiques dels textos especialitzats; i és als anys noranta amb la creació de grans corpus textuals informatitzats que els primers programes d'extracció *semiautomàtica* [3]de terminologia comencen a donar resultats positius.

Certament durant aquesta darrera dècada, lingüistes computacionals, lingüistes aplicats, mediadors lingüístics (traductors, terminòlegs, intèrprets, periodistes científics, etc.), informàtics, enginyers, documentalistes, s'han interessat per motius ben diferents en la possibilitat d'aconseguir aïllar informàticament terminologia a partir de textos.

Les finalitats que empenyen aquests distints col·lectius professionals a dissenyar eines informàtiques que puguin extreure la terminologia directament dels textos són molt diverses, entre les quals podem destacar:

- la creació de glossaris, vocabularis i diccionaris terminològics
- la creació de programes de traducció automàtica
- la indexació de textos
- la creació de bases de coneixement
- la creació de sistemes d'hipertext
- la creació de sistemes experts
- l'anàlisi de corpus

De tots aquests objectius, el primer, *la creació de glossaris, vocabularis i diccionaris terminològics,* és el que ha donat lloc a més sistemes de detecció automàtica de terminologia. Com és sabut, la fase clau --i segurament que també la més llarga-- de la creació d'un glossari terminològic és la del **buidatge** de les unitats terminològiques a partir de textos especialitzats. Aquesta fase comprèn tres etapes:

1. el reconeixement dels termes, simples i complexos
2. la delimitació dels termes complexos
3. la determinació de la pertinença del termes al camp conceptual en què s'està treballant

Aquestes tres tasques fins ara es realitzaven manualment, i podien resultar llargues, avorrides, repetitives i, sobretot, corrien el risc de ser asistemàtiques i subjectives. Amb l'automatització total o parcial d'aquests processos s'aconsegueix guanyar velocitat, sistematicitat i objectivitat en l'elaboració de glossaris. Així doncs, els sistemes de detecció automàtica de terminologia neixen com una eina d'ajuda al treball terminològic.

[2]Voldríem agrair al Dr. Horacio Rodríguez (Universitat Politècnica de Catalunya), al Dr. Christian Jacquemin (Université de Nantes), al Dr. Lluís de Yzaguirre (Universitat Pompeu Fabra) i a la professora Judit Freixa (Universitat Pompeu Fabra) la lectura minuciosa de versions preliminars d'aquest text.

[3]En aquest treball hem utilitzat indistintament les paraules *extractor* i *detector* amb la finalitat de donar una visió més global d'aquest tipus d'eines. Tot i que som concients que alguns autors distingeixen clarament aquests dos termes.

Des de l'aparició de TERMINO l'any 1990 fins avui dia s'han dut a terme una sèrie de projectes per dissenyar diferents tipus de detectors automàtics de terminologia, però malgrat la gran quantitat d'estudis que s'estan realitzant en aquesta línia, l'automatització de la fase de buidatge terminològic encara no està totalment resolta.

Els principals problemes amb què es troben actualment els extractors de termes són:

1. la delimitació dels termes complexos, és a dir, saber on comença i on acaba un sintagma terminològic,
2. el reconeixement dels termes complexos, és a dir, decidir si un segment discursiu és un sintagma terminològic o és un segment lliure,
3. la identificació del caràcter terminològic d'una unitat lèxica, és a dir, saber quan en un text especialitzat una unitat lèxica és terminològica i quan pertany a la llengua general, i
4. la pertinença d'una unitat terminològica per a un vocabulari concret (aspecte poc estudiat encara des del punt de vista de l'automatització)

Totes aquestes dificultats responen a dues qüestions teòriques de base no resoltes per la terminologia: Què és una unitat terminològica? i Què és una àrea d'especialitat? La distinció amb precisió entre una unitat lèxica, un frasema i una combinació de mots lliure i la distinció entre una unitat lèxica de la llengua comuna i una unitat lèxica pròpia d'una àrea d'especialitat concreta, són dos dels conceptes que necessiten estar predefinits per poder dissenyar un detector de terminologia.

Consegüentment, tots els mètodes que fins ara s'han creat són **semiautomàtics.** Aquest fet comporta que al final del procés de buidatge s'arribi no a una llista de termes --com seria desitjable-, sinó a una llista de *candidats a terme*, i és sempre la competència cognitiva de l'expert i l'experiència de l'ús qui en últim lloc ha d'acabar de decidir quins d'aquests candidats són termes. En definitiva, no tenim encara eines informàtiques, ni de base estadística ni de base lingüística, que permetin detectar, delimitar i desambiguar totes les unitats terminològiques a partir de textos informatitzats.

Com ja hem avançat, els sistemes d'extracció semiautomàtica de terminologia es basen en diversos tipus de coneixement:

1. coneixement lingüístic
2. coneixement estadístic
3. coneixement híbrid: estadístic i lingüístic

Hi ha, doncs, diferents aproximacions a la detecció automàtica de termes d'un text. Bàsicament tots els sistemes analitzen un corpus de text d'especialitat en suport electrònic del qual s'extreuen llistes de seqüències de paraules (*candidats a terme*) que el terminòleg ha de validar. Per facilitar la feina del terminòleg se sol acompanyar el *candidat a terme* del context d'aparició i, eventualment, d'alguna informació addicional (freqüència d'aparició, relació amb altres termes, etc.).
Paral·lelament a la detecció de termes, trobem la indexació automàtica de documents (*information retrieval*: IR). Aquest camp d'aplicació de tècniques de processament de llenguatge natural té punts

en comú importants amb la detecció automàtica de termes ja que les seqüències de paraules que serveixen per indexar un document normalment són també unitats terminològiques.

Aquesta coincidència d'objectius justifica que molts dels sistemes d'extracció que s'exposaran més endavant provenen de la IR així com l'anàlisi d'algun sistema específic d'IR sense cap aplicació en l'extracció de termes.

La diferència entre aquestes dues aproximacions radica en el fet que una eina de detecció de termes hauria d'extreure *totes* les unitats terminològiques d'un text, mentre que en IR l'important és extreure només les paraules o seqüències de paraules que descriuen millor el contingut del document, independentment de les seves característiques gramaticals.

L'aproximació típica de la IR és processar els documents per extreure els anomenats *termes d'indexació*. Aquests termes normalment són paraules aïllades amb un pes semàntic que proporciona informació sobre la seva *bondat*[4]en descriure el document. La interrogació es processa d'una manera similar per extreure els *termes d'interrogació*. La rellevància d'un document en relació a una interrogació es basa exclusivament en els termes que el representen i, per tant, la seva elecció es crítica.

Sovint, però, aquests termes d'indexació són paraules simples, malgrat que se sap que les paraules aïllades rarament són prou significatives per decidir amb seguretat el valor semàntic d'un document en relació a una interrogació. Aquest fet ha donat lloc a l'aparició progressiva, en les avaluacions TREC[5], de sistemes que indexen per paraules i per seqüències de paraules utilitzant tècniques pròpies del processament del llenguatge natural.

Els sistemes que utilitzen una aproximació estadística basen el seu funcionament en la detecció de dues o més unitats lèxiques concorren freqüentment per sobre d'un cert llindar. Es considera que aquesta situació no és casual, sinó que correspon a un ús particular d'aquestes unitats lèxiques. Aquest mateix principi, anomenat *informació mútua* (*mutual information*), s'utilitza també en altres àrees de la ciència com les telecomunicacions o la física. Els detectors de termes solen usar aquest principi com a filtre previ a un processament que utilitzi coneixement lingüístic.

El problema associat a aquest tipus d'aproximació és el fet que els termes amb freqüència d'aparició molt baixa tenen tendència a escapar-se del sistema d'extracció. És important remarcar que aquests sistemes utilitzen bàsicament informació numèrica i, per això, són relativament independents d'una llengua determinada.
Les dues mesures més utilitzades en la valoració d'aquests sistemes provenen de la IR: *recall* i *precisió*. El *recall* es defineix com la relació entre el total de termes recuperats i el total de termes existent en el document que s'està explorant. La *precisió*, en canvi, és la relació entre els candidats a termes extrets que realment són termes i el total de candidats a termes trobats. Aquestes mesures

[4] La *bondat* és una mesura del valor d'un índex expressat com un valor numèric que es calcula bàsicament a partir de la freqüència d'aparició. En el càlcul no es tenen en compte les paraules gramaticals.
[5] TREC (*Text Retrieval Engineering Conference*) és el nom amb què es coneixen una sèrie de conferències patrocinades per les agències NIST (*National Institute of Standards and Technology*) i *DARPA (Defense Advanced Research Projects Agency*) dels EEUU. L'objectiu d'aquest fòrum és fomentar la recerca sobre el tractament de grans volums d'informació textual. Fins ara s'han realitzat cinc conferències d'aquest tipus, per a més informació podeu consultar la URL: *http://www-nlpir.nist.gov/TREC*.

es poden interpretar com la capacitat del sistema de detecció per extreure tots els termes presents en un document (*recall*) i la capacitat de discriminació entre les unitats detectades pel sistema que són termes i les que no ho són (*precisió*). El fet que el *recall* tingui en compte **tots** els termes presents en el document implica que sigui a la vegada una xifra molt més difícil de calcular i de millorar que la *precisió*.

En contraposició a aquesta aproximació més tradicional, n'han sorgit d'altres que intenten resoldre el problema aprofitant el coneixement lingüístic. Aquest coneixement pot ser informació:

- específica dels termes: amb la detecció de les seqüències de categories típiques de les unitats terminològiques complexes, com, per exemple, les seqüències Nom-Adjectiu, Nom-Preposició-Nom. Per la qual cosa s'ha de recórrer a les *expressions regulars* i tècniques d'*autòmats d'estats finits*[6].

- genèrica del llenguatge: amb la utilització de sistemes de processament de llenguatge natural més complexos que treballen a partir de la detecció d'estructures lingüístiques més bàsiques (sintagma nominal, sintagma preposicional, etc.).

En ambdues aproximacions es comença associant a cada paraula una categoria morfològica. Per realitzar aquesta tasca, es proposen alternatives molt diverses: des de sistemes molt bastos que no utilitzen cap diccionari fins a sistemes complexos amb una anàlisi morfològica molt detallada i una etapa final de desambiguació.

El sistemes que utilitzen informació estructural recorren a tècniques d'anàlisi parcial per detectar les estructures sintagmàtiques potencialment terminològiques. També trobem sistemes que aprofitant el coneixement del que no és un terme treballen a mig camí entre els sistemes ja mencionats. També hi ha d'altres que basen el seu funcionament en reutilitzar les bases de dades terminològiques existents per trobar termes, variants o nous termes.

Tots els sistemes basats en coneixement lingüístic solen utilitzar com a unitats de mesura de l'eficàcia el *soroll* i el *silenci*[7]. En el primer cas, es tracta d'avaluar la proporció entre els candidats rebutjats i els acceptats. En el segon cas, s'avalua els termes presents en el text analitzat, però no detectats pel sistema. En tots dos casos las mesures s'expressen en forma de percentatge. El soroll és un problema típic dels sistemes que utilitzen aquesta aproximació. També els errors en l'assignació de la categoria morfològica són un problema recurrent en aquest sistemes.
El tipus de coneixement utilitzat fa que els sistemes d'aquest tipus siguin propis de cada llengua i, per tant, la utilització en una llengua diferent necessita d'un estudi lingüístic previ i probablement del redisseny de moltes parts del sistema.

[6] Un autòmat d'estats finits està representat pel conjunt: {A, E, I, F, T}, on A és l'alfabet, E el conjunt d'estats, I i F són els conjunts d'estats inicials i finals i T el conjunt de transicions possibles. El procés de reconeixement d'un patró morfològic consisteix a situar-se en un del nodes inicials i desplaçar-se dintre de l'autòmat tot seguint les transicions indicades per l'etiqueta corrent. Si s'arriba a un dels estats finals es diu que el patró pertany al llenguatge representat per l'autòmat i la seqüència es dóna per reconeguda.
[7] Els termes *recall-silenci* i *precisicó-soroll* dónen la mateixa informació tenint en compte, però, punts de vista diferents. Així, per exemple, un sistema amb un *recall* del 75% li correspondria una xifra de *silenci* del 25%.

La intel·ligència artificial ha obtingut tradicionalment el coneixement directament dels experts en cada àrea d'interès. Aquesta aproximació s'ha trobat amb una sèrie de dificultats que ha portat alguns investigadors a considerar en l'automatització i en la sistematització de l'adquisició del coneixement els avantatges d'una aproximació terminològica. Així, alguns investigadors (Condamines, 1995) han proposat la creació de bases de coneixements terminològics per tal d'incloure coneixement lingüístic en les bases de dades tradicionals. Aquesta aproximació és molt nova i no hi ha encara cap base de dades d'aquestes característiques que pugui ser utilitzada en l'extracció de termes, l'ús d'informació semàntica és ara mateix molt escassa. En aquest sentit, s'ha començat a treballar amb llistes tancades de les paraules amb menys pes semàntic dintre d'una àrea de especialitat molt concreta.

En aquest escrit, presentem una anàlisi dels principals sistemes d'extracció de terminologia amb la finalitat de dibuixar un panorama ampli de l'estat de la qüestió, i d'aquesta manera disposar d'elements de judici per millorar aquests programes.

Hem dividit l'informe en dos grans apartats; el primer, molt més extens, és una descripció de diversos sistemes d'extracció de terminologia acompanyada d'una breu valoració en què hem tingut en compte tant els punts febles com els aspectes més interessants.

En el segon apartat, hem classificat els sistemes d'extracció de terminologia analitzats a partir dels set paràmetres següents:

- els nivells d'informació que utilitzen
- les estratègies que fan servir per delimitar els termes
- els recursos que usen per filtrar els termes
- les estratègies d'adquisició de coneixement
- el sistema de presentació dels termes reconeguts
- el sistema d'interacció amb l'usuari
- els resultats que obtenen (és a dir, el percentatge de silenci i el percentatge de soroll).

L'informe finalitza amb unes conclusions sobre els sistemes analitzats i sobre els criteris que podrien guiar una proposta integradora.

Els sistemes de detecció semiautomàtica de terminologia analitzats són els següents[8]:

1. ANA
2. ATELIER/FX
3. Autolex
4. Blank
5. CLARIT
6. Daille
7. Drouin
8. FASTR
9. Heid
10. LEXTER
11. NEURAL
12. NODALIDA-95
13. SBIC
14. Termight
15. TERMINO
16. TERMS
17. STELLA

[8] Els sistemes analitzats es referencien sempre que és possible pel seu nom, i només quan no en tenen públicament s'ha utilitzat el nom de l'autor. En tots els casos s'ha respectat la grafia utilitzada pels autors.

2. Descripció dels principals sistemes d'extracció de terminologia

En aquest apartat presentem una descripció crítica de disset sistemes d'extracció semiautomàtica de terminologia. En tots els casos, presentem les informacions següents:

1. Les dades de referència del sistema, és a dir l'autor, la publicació on es ressenya per primer cop l'eina, l'objectiu principal del sistema i les llengües per a les quals ha estat dissenyada l'aplicació.
2. Un resum de les característiques principals del sistema.
3. Una valoració breu dels aspectes més rellevants.

2.1 ANA[9]

Autors**:** Enguehard C.; Pantera, P.
Publicació**:** Automatic Natural Acquisition of a Terminology
Data**:** 1994
Objectiu**:** Eina d'extracció de candidats a terme
Llengües de treball**:** Qualsevol (aplicat al francès i a l'anglès)

Sinopsi

Aquest sistema d'extracció (ANA) parteix dels tres principis següents:

- no utilitzar coneixement lingüístic
- tractar tant amb textos escrits com orals (transcripcions d'entrevistes)
- ser insensible a errors de sintaxi

Seguint la tendència actual d'utilitzar tècniques estadístiques en els estudis sobre llenguatge natural, els autors fan servir la informació mútua com a mesura d'associació lèxica[10]. Per no haver de recórrer al coneixement lingüístic es crea el concepte de "reconeixement flexible de cadenes", que defineix una funció matemàtica per determinar el grau de similitud entre dues paraules i no haver de recórrer a cap eina d'anàlisi morfològica. Per exemple, la cadena *colour of painting* representa també aquestes altres: *colours of paintings, colour of this painting*, *colour of any painting*, etc. El sistema no disposa de diccionari ni de gramàtica.

L'arquitectura general d'ANA conté dos mòduls bàsics:

- mòdul de familiarització
- mòdul de descobriment

[9] Automatic Natural Acquisition
[10] Exemples significatius d'utilització d'aquestes tècniques són els treballs de Church/Hanks, 1989 (sobre associació de paraules) i Smadja, 1991 (sobre extracció de col·locacions a partir de corpus molt extensos).

El primer mòdul serveix per determinar els tres grups de paraules que es defineixen a continuació i que representen l'únic coneixement necessari per a la detecció:

- paraules gramaticals o *function words*: a, any, for, in, is, of, to, ...
- paraules que serveixen per establir una relació semàntica o *scheme words*, p. e. en "*box of nails*", la paraula *of* indica una certa relació entre *box* i *nails*.
- el conjunt de termes que constitueix el nucli del sistema i punt de partida per a la detecció o *bootstrap*.

El segon mòdul, el de descobriment, està constituït bàsicament per un procediment d'adquisició incremental de nous termes a partir dels ja coneguts. A més, es generen automàticament enllaços entre els termes detectats que serviran per construir una xarxa semàntica. La concurrència de paraules, en la qual es basa l'etapa de descobriment, pot tenir interpretacions de tres tipus:

- *expressió*: reben aquesta qualificació els termes ja existents que apareixen sovint (sostre T_{EXP}) en la mateixa finestra. La nova paraula és considerada un nou terme i inclosa en la xarxa semàntica. Per exemple, si el sistema troba seqüències d'aquest tipus:

  ... the *diesel engine* is...
  ... this *diesel engine* has ...
  ... a *diesel engine* never ...

  la seqüència *diesel engine* és acceptada com un nou terme i s'incorpora a la xarxa semàntica com un nou nus amb enllaços cap a *diesel* i cap a *engine*

- *candidat*: quan un terme ja reconegut apareix sovint (sostre T_{CAND}) amb una altra paraula i una *scheme word,*

  ... any *shade* of wood ...
  ... this *shade* of colour ...
  ... the *shade* of the bench ...

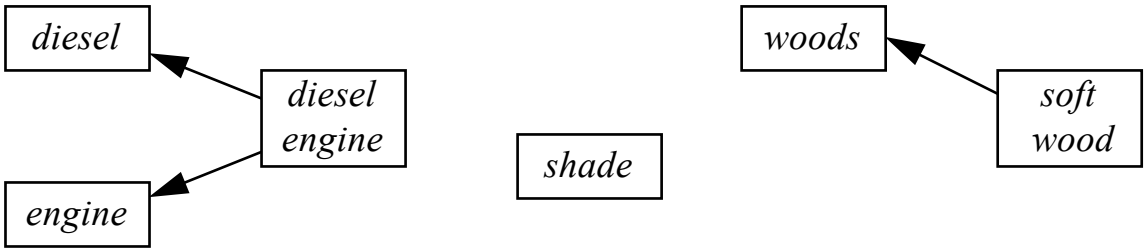
  en aquest cas *shade* es converteix en un terme nou i en un nou nus de la xarxa semàntica.

- *expansió*: un terme ja existent apareix sovint (sostre T_{EXPA}) en la mateixa seqüència de paraules sense incloure cap *scheme word*,

  ... use any *soft* woods to ...
  ... this *soft* woods or ...
  .. cheapest *soft* woods comes ...

  com a resultat *soft wood* s'incorpora a la llista de termes i a la xarxa semàntica com un nou nus amb un enllaç a *woods.*

En la figura es mostra els nusos i les relacions de la xarxa semàntica com a resultat de la detecció dels termes dels exemples anteriors:



El sistema continua buscant recursivament elements amb les tres interpretacions ja esmentades fins que no troba cap altre terme. En les proves realitzades pels mateixos autors el sistema ha processat un document en anglès amb 25.000 paraules i 29 termes de referència. En aquestes condicions, el sistema extreu 200 termes nous amb una taxa d'error de l'ordre del 25%.

Valoració

Un aspecte negatiu d'aquest sistema és que les unitats terminològiques que s'afegeixen a la llista de termes vàlids després de cada cicle no són validades; això implica que aquest tractament permet d'afegir termes no vàlids que incrementen la llista de termes.

En canvi, és molt valuosa la idea de minimització dels recursos lingüístics, encara que no s'aporten dades sobre l'eficàcia d'aquesta proposta.

2.2 Atelier FX

Autors**:** Equip d'ATO
Publicació**:** web ATO[11]
Data**:** 1989
Objectiu**:** Analitzar el contingut de textos escrits en llenguatge natural.
Llengua de treball**:** Francès

Sinopsi

[11]URL: *http://www.ling.uqam.ca/Ato/FX/AtelierFX.html*

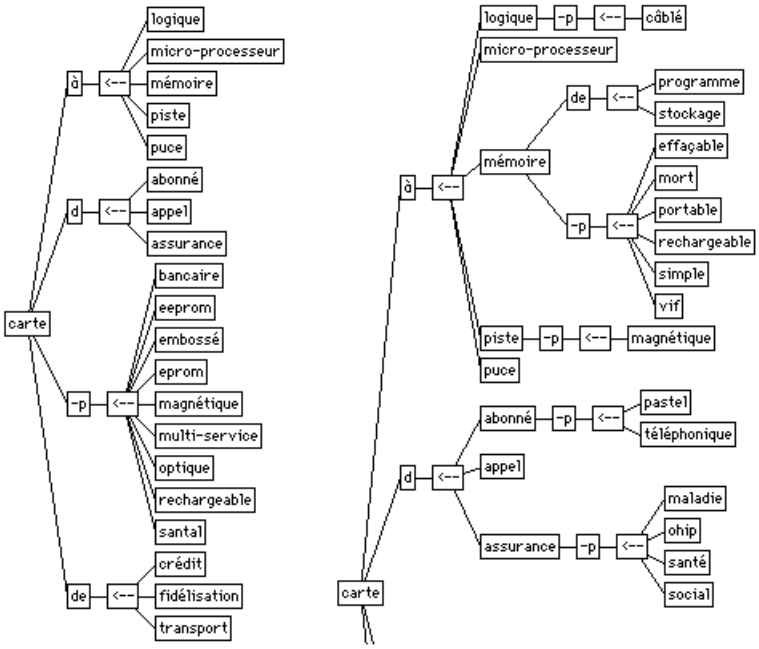
Atelier FX està compost de quatre mòduls:

1. El **llenguatge FX**, que permet utilitzar una tècnica de programació que els autors denominen *en faisceaux*.
2. El **càlcul de *saillance* FXS**, que permet la comparació d'objectes construïts per un programa FX
3. El **mòdul NOMINO**, que és un analitzador morfològic per al francès
4. El **programa FX-Fiches**, que permet construir bases de coneixements a partir dels tres mòduls anteriors.

Pel que fa a l'extracció de termes ens interessa examinar bàsicament el tercer mòdul d'Atelier FX anomenat NOMINO.

NOMINO és un analitzador morfològic que permet lematitzar les formes lèxiques i donar per a cada un dels lemes la categoria morfològica desambiguada corresponent. Els seus autors consideren les unitats complexes nominals (UCN) com una categoria gramatical --al costat del nom i de l'adjectiu--, que defineixen com un grup de mots construïts a l'entorn d'un nucli nominal. Aquestes UCN són bones candidates a esdevenir unitats referencials, és a dir, unitats terminològiques i es poden desglossar en nucli i expansió.

Una de les altres possibilitats que ofereix el sistema és la construcció, totalment automàtica, de xarxes d'UCN. Un exemple d'aquestes xarxes és la següent, extreta a partir del mots que componen la ucn *carta memóire*:



Segons els autors, aquestes xarxes d'UCN representen esquemes semàntics que permeten visualitzar ràpidament el contingut del text. A partir d'aquestes xarxes d'UCN i de les descripcions

lingüístiques obtingudes per NOMINO es poden extreure bases de coneixements amb una presentació hipertextual.

Una altra aplicació d'Atelier és la possibilitat d'obtenir llistes dels nuclis d'ucn que reben més extensions, o a la inversa, llistes de les expansió més productives.

Valoració

Destaquem les observacions següents:

- Atelier FX no és un sistema dissenyat per extreure unitats terminològiques, sinó per afavorir la lectura d'un document i poder fer diagnòstics. Es tracta, per tant, d'un paquet de programes que es podria incloure en els anomenats sistemes experts.
- Per a l'extracció de terminologia és especialment interessant l'aplicació que permet establir automàticament xarxes d'unitats complexes nominals. Aquestes xarxes, a parer nostre, representen un primer pas cap a l'anàlisi semàntica de corpus.
- També és molt interessant el fet de poder extreure de manera automàtica llistes de les extensions més freqüents i llistes dels nuclis més repetits, perquè a partir d'aquestes llistes segurament es poden obtenir dades per establir relacions semàntiques entre les unitats terminològiques.

2.3 Autolex

Autor**:**	Planas, A.
Publicació**:**	*AUTOLEX: Sistema para la gestión de bases de datos terminológicas. Herramienta para la traducción asistida por computadora.* Ciencias de la información Vol. 25, No.25.
Data**:**	1994
Objectiu**:**	Creació i manteniment de diccionaris especialitzats
Llengua de treball**:**	Castellà

Sinopsi

Autolex és una eina informàtica que no té com a finalitat última la detecció de termes encara que el resultat que s'obté pot ser acceptable en determinats casos. El sistema incorpora facilitats per a la gestió dels termes obtinguts i la seva traducció a altres llengües (aquestes dues aplicacions, però, no es comentaran perquè l'objectiu d'aquest estudi és un altre).

La característica principal d'aquest sistema és que treballa sense fer ús de cap informació lingüística, tret de la llista de paraules buides que actuen de frontera de terme. Per exemple, una possible llista de paraules buides per al català a l'inici d'un terme podria ser la següent: *un, una, uns, unes, el, la, els, les, del*, etc. El mecanisme de funcionament és molt similar al sistema SBIC ja que tots dos sistemes deixen al terminòleg/especialista la tasca de seleccionar manualment els

termes pertinents. Malgrat aquesta similitud, s'ha de remarcar que el sistema Autolex és relativament més ràpid que el SBIC.

Valoració

La manca d'utilització de coneixement lingüístic, la necessitat d'una selecció i validació manual gairebé de la totalitat de seqüències i l'ús escàs que n'han fet altres equips ens porta a considerar aquest sistema com a poc profitós per a l'extracció automàtica de terminologia.

De totes maneres, el fet que aquesta eina estigui disponible per a microordinadors fa pensar en la seva utilització en l'àmbit de l'ensenyament o bé en la realització de tasques puntuals no excessivament extenses.

2.4 Blank

Autora**:** Blank, I.
Publicació**:** *Méthodes pour l'extraction de terminologie bilingue*
Data**:** 1995
Objectiu**:** Extracció de terminologia bilingüe en textos tècnics
Llengües de treball**:** Alemany, anglès i francès

Sinopsi

Per extreure terminologia en diferents llengües simultàniament Blank utilitza un conjunt de mètodes de base estadística. L'extracció de terminologia segueix quatre fases:

a) selecció del corpus apropiat
b) pretractament del corpus
c) extracció de termes monolingües
d) alineació a nivell de frases de textos en diferents llengües

Per a cada una de les fases s'usen uns programes determinats --alguns ja existents en el mercat i, per tant, no creats per a l'ocasió.

Fase 1. El corpus

L'equip de Blank treballa amb un corpus trilingüe (alemany-anglès-francès) de 12 milions de mots que corresponen a 1000 decisions de la Cambra de Recurs de l'Oficina Europea de Patents (OBE). Aquest corpus es caracteritza bàsicament per 4 aspectes:

- És un corpus **textual;** comprèn 1000 documents.
- És un corpus **multilingüe;** els textos són paral·lels en tres llengües.
- És un corpus **extens;** es garanteix l'operacionalitat de l'aplicació de mètodes estadístics.
- És un corpus **especialitzat**; els textos tenen valor jurídic.

Fase 2. El pretractament

En aquesta fase, es neteja i s'estructura el corpus. A través de ***programes d'identificació de finals de frases*** per a cada llengua, es marquen dins de cada text els paràgrafs i les frases. Aquests programes són eficients en un 95% del casos.

Fase 3. Extracció de termes

Mitjançant el programa INTEX (Silbertein, 1993) —eina basada en diccionaris i gramàtiques útils per a l'anàlisi lingüística de textos — es localitzen els patrons sintàctics. A continuació, autòmats finits identifiquen aquests patrons en els textos pretractats. Els termes que s'obtenen amb aquests patrons es classifiquen segons la seva freqüència. Finalment, un especialista els avalua. A l'entorn del 80% dels candidats a terme amb un nombre d'ocurrències superior a 3 són realment termes.

Fase 4. Alineament de frases

L'alineament de les frases es pot aconseguir per diferents aproximacions:

a) mètodes purament estadístics
b) mètodes que combinen les informacions lingüístiques i les estadístiques

L'autora ha aplicat sobre un corpus de 400.000 mots un mètode estadístic i un mètode lingüístic per tal de valorar-ne els avantatges i els inconvenients.

El mètode de Gale i Church (1991), de naturalesa estadística, es basa únicament en la llargada de les frases. El resultat és que alinea el 95 % de les frases de la mostra.

En canvi, el mètode de Ray i Röescheisen (1988), que és més lingüístic, es basa en l'alineament de les paraules que contenen les frases a través d'un procés algorítmic de relaxació. Tenint en compte punts d'anclatge com el començament i el final d'un text, l'algoritme divideix un paràgraf en segments susceptibles de correspondre's. Després es comparen els mots de dos en dos i es calcula el seu grau de similitud; aquests mots serveixen alhora com a punts d'anclatge en la següent divisió del text. L'algoritme continua fins que no troba cap punt nou d'anclatge. El resultat és que alinea del 75% al 90% de les frases de la mostra. Al final de l'execució d'aquest programa s'extreu en dues taules: una amb les frases alineades i una altra amb els mots alineats.

Amb aquest sistema, però, només un terç dels termes reconeguts anteriorment són alineats amb la seva traducció corresponent. D'aquest terç d'unitats terminològiques, un 85% dels termes complexos i un 75% dels termes simples són alienats i traduïts correctament.

Aquest algoritme és satisfactori pel que fa a la qualitat, però no a la quantitat. Es pretén augmentar la mostra amb textos més llargs i de matèries similars per millorar els resultats.

Valoració

Destaquem les observacions següents:

- Aquest sistema només extreu unitats terminològiques nominals.
- No tenim dades de la quantitat de silenci que el sistema produeix.
- En canvi, sabem que produeix un gran percentatge de soroll (20%).
- Parteix d'un nombre molt reduït de patrons sintàctics determinats, per tant no pot recollir unitats neològiques des del punt de vista de l'estructura sintàctica o unitats amb patrons poc freqüents.
- L'alineació dels termes està encara molt poc treballada i els èxits són baixos.
- El sistema es basa en disposar de documents paral·lels en tres llengües, cosa que és poc difícil de recopilar[12].

2.5 CLARIT[13]

Autors:	Evans, D. A.; Zhai, C.
Publicació:	*Noun-phrase Analysis in Unrestricted Text for information retrieval*.
Data:	1996
Objectiu:	Indexació de documents
Llengua de treball:	Anglès

La tasca d'indexació de documents (IR: *information retrieval*) és un camp important d'aplicació de les tècniques de processament de llenguatge natural. Aquesta branca té punts en comú importants amb la detecció de termes ja que les seqüències de paraules que serveixen per indexar un document normalment són també unitats terminològiques. La diferència radica en el fet que una eina d'extracció de termes hauria d'extreure totes les unitats terminològiques d'un text, mentre que per a la IR només són importants les paraules o seqüències de paraules que descriuen millor el contingut del document independentment de les característiques gramaticals que tinguin.

CLARIT s'inscriu entre el grup de sistemes que es decanten per un processament elaborat dels textos per tal de detectar termes complexos que puguin servir per a una caracterització més

[12]Normalment quan es parla de corpus paral·lels es fa referència al domini del dret, en altres casos es parla no de corpus paral·lels, sinó de corpus comparables. Un exemple, però, de corpus paral.lel en més d'un domini seria el Corpus de l'IULA que treballa en cinc àrees temàtiques prioritzant aquest tipus de document Bach, C. i al., 1997.

[13]CLARIT és el nom d'un producte d'informàtica documental desenvolupat per CLARITECH Corp. Per a més informació podeu consultar la URL: *http://www.clarit.com*.

adequada dels documents[14]. És per aquest motiu que hem inclòs aquest sistema entre els detectors de terminologia.

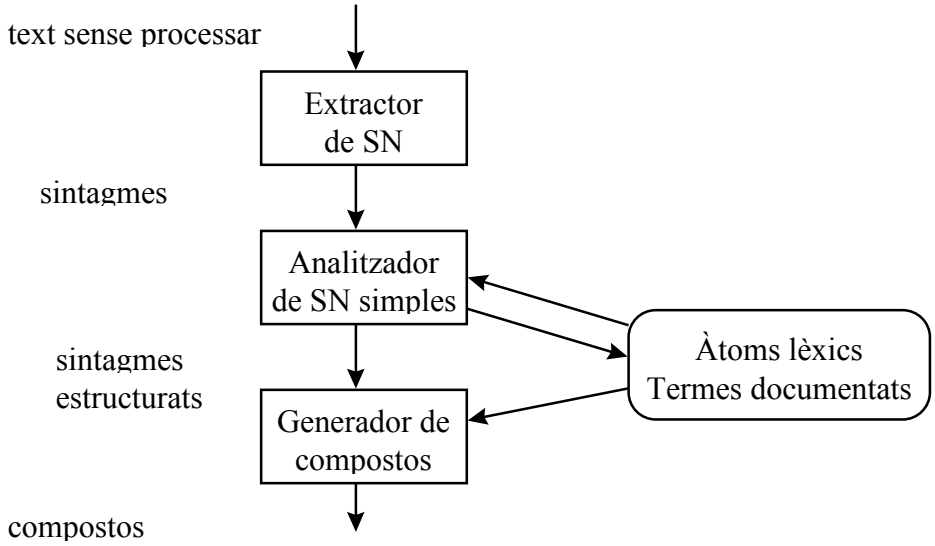
Els autors proposen utilitzar els següents tipus de termes per a la indexació :

- àtoms lèxics (*hot dog*, *stainless steel*, *data base*, *on line*, ...)
- parelles nucli-modificador (*treated strip*, ...)
- subcompostos (*stainless steel strip*, ...)
- modificadors lligats amb preposició (*quality surface* vs *quality of surface*)

La metodologia proposada comença amb l'anàlisi morfològica de les paraules i la detecció de frases nominals. El sistema distingeix entre frases nominals simples i frases nominals complexes. Les frases simples corresponen als termes d'indexació, mentre que les complexes són frases simples concatenades amb preposicions.

La idea subjacent és aplicar l'estadística a la lingüística de corpus. L'estadística utilitzada es centra directament en els documents, és a dir, no hi ha cap corpus d'entrenament previ. Els coneixements lingüístics permeten agilitar el càlcul eliminant estructures irrellevants, millorar la fiabilitat de les decisions estadístiques i ajustar els paràmetres estadístics.

El procés que se segueix està representat en la figura següent :



En primer lloc, s'analitza el text en brut per extreure'n els sintagmes nominals. A continuació, cada sintagma nominal s'analitza en un procés recursiu per tal de trobar els agrupaments més segurs. En aquesta fase es detecten també els àtoms lèxics i s'estructuren els sintagmes nominals. Finalment, s'arriba a la fase de generació on s'obté la resta de compostos.

Per a la detecció dels àtoms lèxics es recorre a les dues regles heurístiques següents:

[14]Vegeu capítol d'introducció, o Hull i al., 1996 per a una altra aproximació similar a la qüestió.

- les paraules que constitueixen un àtom lèxic mantenen una relació molt forta i tenen tendència a aparèixer lexicalitzades com una única unitat lèxica
- quan un àtom lèxic funciona de sintagma nominal mai o quasi mai admet paraules inserides.

La primera condició es realitza comprovant que la freqüència d'una parella P_1P_2 que es vol agrupar és més alta que la de qualsevol altra parella (que inclogui P_1 o P_2) del sintagma nominal que s'està processant.

Per a la segona condició es comparen les freqüències d'aparició conjunta i separada establint un sostre per sota del qual l'associació és rebutjada. Aquest sostre és variable en funció de la categoria morfològica de la seqüència. En els textos anglesos, la seqüència més afavorida és nom-nom.

L'anàlisi del sintagmes nominals també és un procés recursiu. En cada nova fase d'anàlisi s'utilitzen els àtoms lèxics més recents per trobar noves associacions que s'utilitzaran en la fase següent. El procés continua fins que s'ha analitzat tot el sintagma nominal. Com a exemple es presenta la seqüència d'anàlisi per a una frase del corpus

general purpose high perfomance computer
general purpose ***[high perfomance]*** *computer*
[general purpose] *[high perfomance] computer*
[general purpose] ***[[high perfomance] computer]***
[[general purpose] [[high perfomance] computer]]

L'ordre en què s'han fet les agrupacions reflecteix quines són les seqüències que tenen un grau d'associació més fiable. Per determinar el grau d'associació es tenen en compte algunes regles :

- els àtoms lèxics tenen la prioritat més alta, igual que les combinacions d'adverbis amb adjectius, participis i verbs progressius
- les parelles impossibles sintàcticament tenen la prioritat més baixa (nom-adj, nom-adv, adj-adj, ...)
- per a la resta de parelles s'utilitza una fórmula que té en compte les freqüències de cada paraula, el grau d'associació d'aquesta paraula amb les altres paraules del sintagma nominal i de dos paràmetres arbitraris

Per evitar els problemes de la poca fiabilitat en considerar la freqüència com el factor més important en el càlcul de l'associació, el grau d'associació entre dues paraules o *bigrams*[15] es recalcula després de cada assignació d'associació.

El sistema s'ha provat fent la tasca real d'indexació de documents i substituint el mòdul estàndard del sistema CLARIT amb un altre que realitzi l'anàlisi que hem esmentat. El corpus de prova i les interrogacions van ser els estàndards utilitzats en les conferències TREC. S'han observat millores tant en el *recall* com en la precisió que, en opinió dels autors, justifica la utilització d'aquestes tècniques. Posteriorment, en el cos del TREC-5 s'ha fet una avaluació més detallada del sistema

[15] Bigram és un model de representació del llenguatge on es fa l'aproximació que la probabilitat d'una paraula només depèn de la paraula precedent:

$$p(s) = p(w_1)\, p(w_2|w_1)\, p(w_3|\, w_1w_2) \ldots p(w_n|w_1\, w_2 \ldots w_{n-1}) = \text{☺}\, p(w_n|w_1\, w_2 \ldots w_{n-1}) \sim \text{☺}\, p(w_n|w_{n-1})$$

(Zhai et al. 1996). La conclusió general és que l'ús d'aquestes tècniques és efectiva, conclusió que reforça la idea de similitud entre els problemes d'indexació de termes i l'extracció de terminologia:

> *A brief manual inspection of the identified lexical atoms shows that many, in fact, most, are not true lexical atoms. However, they are usually good terminological phrases such as "stock market" and "annual meeting".*[16]

Valoració

El sistema sembla molt interessant i en principi l'aplicació d'algunes idees de base a la detecció de terminologia sembla factible; de fet, el sistema té semblances amb aspectes de la proposta de Daille 94 (estadística guiada per coneixement lingüístic).

Cal tenir en compte que, encara que els problemes d'extracció de terminologia i d'indexació de documents són semblants, no són idèntics i, per tant, moltes decisions s'haurien de reconsiderar des del punt de vista estricte de la detecció de termes.

També cal remarcar que aquest sistema, per definició, només extreu unitats terminològiques nominals i que les dades que es donen del comportament del sistema són en relació a l'aplicació per a la qual s'ha dissenyat.

2.6 Daille-94

Autor**:** Daille, B.
Publicació**:** *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*.
Data**:** 1994
Objectiu**:** Eina d'extracció de candidats a terme
Llengua de treball**:** Francès

La idea bàsica d'aquest treball és combinar el coneixement lingüístic amb mesures estadístiques. En aquest context, el corpus ha d'incorporar, en primer lloc, informació morfològica. A continuació es crea una llista de candidats a terme en base a les seqüències de text que responen a patrons sintàctics de formació de termes. Amb aquesta informació, s'utilitzen mètodes estadístics per filtrar aquesta llista. Aquest processament final es diferencia d'altres treballs que utilitzen només recursos lingüístics.

Les proves s'han realitzat sobre dos corpus de l'àmbit de les telecomunicacions de 500.000 paraules cadascun. Partint del supòsit que tot banc terminològic està format bàsicament per noms compostos, el programa es centra en la detecció dels **noms compostos binaris**, i exclou les altres categories de concurrències. Aquest supòsit es basa en el fet que aquest tipus de noms són els més nombrosos en

[16] Els autors no aclareixen els motius d'aquesta afirmació. A més, a partit dels exemples proposats no es pot deduir els criteris que serveixen per distingir els àtoms lèxics de les frases terminològiques.

els llenguatges d'especialitat. A més a més, la majoria de compostos de longitud igual o major a 3 constituents poden descomposar-se en forma binària.

Els patrons que l'autora ha considerat rellevants per al francès són: N_1 PREP (DET) N_2, i N ADJ juntament amb algunes variants tractades específicament com ara *N_1 PREP de(DET) N_2* i *N_1 PREP à(DET) N_2*, i estructures com ara la coordinació a la dreta o a l'esquerra. Sobre aquests patrons s'apliquen els algoritmes estadístics. L'autora és conscient que l'aplicació de mesures estadístiques implica una certa taxa de silenci, és a dir, que termes de freqüència molt baixa no seran reconeguts.

La tècnica utilitzada per reconèixer els patrons és la dels autòmats finits. Els autòmats són representats per un subconjunt d'etiquetes gramaticals a les quals s'afegeixen alguns lemes, formes flexionades i algun signe de puntuació. Així, podem considerar els autòmats com a filtres lingüístics que seleccionen els patrons definits i en determinen la freqüència d'aparició, distància i variància. Cadascun dels patrons morfosintàctics té associat un autòmat finit específic.

El tractament estadístic que s'aplica al corpus es basa en una sèrie molt extensa de mesures estadístiques que s'agrupen en les classes següents: mesures de freqüència, criteris d'associació, criteris de diversitat i mesures de distància. La idea de partida és considerar els dos lemes que formen una parella dintre d'un patró com a dos variables sobre les quals es vol mesurar el grau de dependència. Les dades es representen en una taula de contingència que té aquest aspecte :

	L_2	L_n
L_1	a	b
L_m	c	d

on : a = ocurrències de la seqüència L_1L_2
b = ocurrències de $L_1 + L_n$ (n != 2)
c = ocurrències de $L_m + L_2$ (m != 1)
d = ocurrències de $L_m + L_n$ (m != 1 i n != 2)

En total s'apliquen 18 mesures diferents amb l'objectiu d'establir el grau d'independència de les variables de la taula de contingència. D'una primera anàlisi dels resultats, l'autora dedueix que només quatre d'aquestes mesures són rellevants per al propòsit fixat:

- la freqüència
- el criteri d'associació amb el numerador al cub[17] (IM^3)
- el criteri de versemblança [18]

[17] Fórmula obtinguda experimentalment per l'autora a partir de la xifra d'associació descrita a Brown i al., 1988 amb l'objectiu d'afavorir les parelles més freqüents:

$$IM^3 = \log_2 (a^3/(a+b)(a-b))$$

En la seqüència de paraules que satisfan el filtre lingüístic N1 (Prep (Det)) N2, els valors més alts depenen del nombre d'ocurrències independentment del nombre de seqüències trobades. Els valors més petits de la xifra d'associació i de la xifra d'associació amb numerador al cub corresponen a paraules que poques vegades apareixen juntes, sinó que solen aparèixer separades (*système terre*, *code signalisation*, ...).

[18] Aquest coeficient introduït per Dunning, 1993 és la prova de relació de versemblança aplicada a una llei binomial.

Loglike = a log a + b log b + c log c + d log d - (a + b) log (a + b)

- el criteri de Fager/MacGowan[19].

Valoració

En aquest sistema, a diferència del que succeeix en altres[20], la freqüència ha resultat ser una de les mesures més importants per a la detecció dels termes d'un àrea, però la classificació que resulta d'aplicar aquesta freqüència selecciona en un nombre important seqüències freqüents que no són termes, i, en canvi, no proposa els termes molt poc freqüents.

La mesura que l'autora pren com a òptima és el criteri de versemblança ja que:

- és un veritable test estadístic
- proposa una classificació que té en compte la freqüència
- té un comportament correcte amb corpus mitjans i grans
- no està definit en els casos que no interessa recollir

De totes maneres, la utilització d'aquesta mesura comporta un cert soroll per raons diverses:

- errors en el marcatge morfològic
- certes combinacions que no són mai compostos (*ko bits*, *à titre d'exemple*, ...)
- combinacions de longitud major o igual a 3 i relacionades amb els problemes de la composició i la modificació (*bande latérale -unique-*, *service fixe -par satellite-*, ...)

2.7 Drouin

Autor**:**	Drouin, P.
Publicació**:**	*Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme*
Data**:**	1995
Objectiu**:**	Presentar un mètode híbrid (anàlisi estadística i anàlisi lingüística) d'identificació automàtica de terminologia
Llengua de treball**:**	Francès

Sinopsi

La proposta de Drouin forma part del conjunt de detectors de terminologia que usen mètodes híbrids: tècniques estadístiques i tècniques lingüístiques. En aquest cas, primer s'apliquen mètodes

(a+c) log (a+c) - (b+d) log (b+d) - (c+d)log (c+d) + N log N

Aquest coeficient selecciona les mateixes parelles que la xifra d'associació amb numerador al cub per als valors més grans i no està definit quan els seus components només apareixen dintre de la parella noms compostos figurats i concurrències de la llengua.

[19] Criteri de biologia i que dóna resultats semblants al criteri d'associació. Dóna molta rellevància a parelles que apareixen sovint juntes i poques vegades separades, sense rebutjar sistemàticament els termes els constituents dels quals apareixen sovint lliurement.

[20] Church/Hanks, 1989.

estadístics als textos per extreure *candidats a termes* i, després, aquests candidats es filtren mitjançant recursos lingüístics.

El sistema està estructurat en les fases següents:

A. Anàlisi estadística
   1. marcatge de candidats a termes
   2. avaluació probabilística de l'estatut dels candidats

B. Anàlisi lingüística
   3. filtratge segons la morfologia dels no-termes complexos
   4. anàlisi de l'autonomia dels candidats
   5. anàlisi dels candidats en context

1. La primera etapa, el marcatge de candidats a termes, consisteix en un algoritme de marcatge molt simple que permet l'obtenció d'encadenaments de mots (*segments*) que es repeteixen "X" vegades. És el mètode presentat per Choueka i al (1983) i Choueka (1988).

2. A la segona etapa, es compara la freqüència d'aparició del segment en relació amb la freqüència d'aparició dels constituents del segment aïllats.

3. La tercera etapa presenta una aproximació diferent: la descripció d'allò que **no pot ser** un terme, usa, per tant, regles negatives. Es construeixen un conjunt de regles per filtrar els candidats rebutjables, com ara les següents:

- "un pronom no pot formar part d'un terme complex"
- "una preposició no pot encapçalar ni acabar un terme complex"
- "un article no pot encapçalar ni acabar un terme complex"
- "una conjunció no pot encapçalar ni acabar un terme complex"
- "un adverbi no pot encapçalar ni acabar un terme complex"
- "un verb conjugat no pot formar part d'un terme"

Observem que les quatre primeres regles es recolzen en categories gramaticals tancades.

4. La quarta fase es basa en la comparació de la freqüència del segment amb la freqüència d'una part d'aquest segment per la qual cosa estableix els dos principis de treball següents:

- si la freqüència d'un segment és la mateixa que la freqüència d'un segment més llarg que el conté, llavors el segment més curt no té un comportament textual autònom;
- si la freqüència d'un segment és superior a la freqüència d'un segment més llarg que el conté, llavors el segment més curt té un comportament textual autònom.

Encara que sigui veritat que la freqüència en terminologia és un indici revelador, aquesta no és de base lingüística, com diuen els autors, sinó que més aviat està en relació amb l'estadística.

5. En l'última etapa, s'acaben de perfilar els candidats mitjançant l'aplicació de cinc regles que tenen a veure amb aspectes gràfics i amb aspectes gramaticals dels termes. Així, podem dir que un segment és més candidat a terme si:

- presenta aspectes gràfics particulars: cometes, majúscules
- va precedit d'un determinant
- va precedit o seguit d'un separador fort
- va seguit d'un verb conjugat
- va seguit d'un pronom relatiu

L'autor afirma que la qualitat dels resultats obtinguts amb aquest mètode híbrid és superior als obtinguts amb mètodes purament lingüístics o purament estadístics.

Valoració

Les observacions més remarcables d'aquest sistema són les següents:

1. L'aplicació en primer lloc d'un mètode estadístic assumeix que hi haurà certs elements terminològics poc freqüents que no es podran recuperar mai i, per tant, el silenci serà un dels punts a tenir en compte quan es valorin els resultats.
2. Només es consideren les unitats terminològiques nominals i es pressuposa que no hi ha verbs complexos en terminologia, si bé és veritat que en els discursos especialitzats la gran majoria de termes són substantius.
3. Només es fa referència al reconeixement de les unitats terminològiques complexes i no es mencionen les estratègies de reconeixement dels termes simples que segueix el programa, és a dir, de quina manera l'ordinador distingeix una unitat lingüística especialitzada d'una unitat lingüística de la llengua general.

2.8 FASTR[21]

Autor**:** Jacquemin, C.
Publicació**:** *A symbolic and surgical acquisition of terms through variation. Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing.*
Data**:** 1996
Objectiu**:** Eina d'extracció de candidats a terme
Llengües de treball**:** Francès, anglès.

El punt de partida d'aquest treball realitzat per C. Jacquemin a la Universitat de Nantes és la idea de construir una eina per a la detecció de terminologia que s'aprofiti dels termes ja coneguts i

[21] FASt Term Recognizer. Projecte desenvolupat a la Universitat de Nantes. Per a més informació podeu consultar la URL següent: *http://perso.wanadoo.fr/christian.jacquemin*

acceptats. Aquests termes poden provenir d'una base de dades existent o ser recollits prèviament per la mateixa eina i validats pel terminòleg. La idea de base és no començar cada vegada de zero.

El primer pas és, doncs, obtenir i analitzar un conjunt de termes ja existents per després extreure'n les variants possibles. Per assolir aquest primer objectiu, utilitza un analitzador parcial que a partir de cada terme obté una regla que després expandeix en les seves variants. Per exemple, tenim el terme *serum albumin* que correspon a una seqüència **Nom-Nom** i que respon a una regla d'aquest tipus:

regla 1: $N_1 \rightarrow N_2\ N_3$
<N2 lema>= serum
<N3 lema>= albumin

Després de l'aplicació d'aquestes regles, actua un conjunt de metaregles a partir de les quals s'obtenen les possibles variants de cada terme de la llista referència. Per posar un exemple, a la regla anterior es podria aplicar la metaregla següent:

$$Coor(X_1 \rightarrow X_2\ X_3) = X_1 \rightarrow X_2\ C_4\ X_5\ X_3$$

i s'obtindria una regla nova:

$$N_1 \rightarrow N_2\ C_4\ X_5\ N_3$$

Mitjançant aquesta nova regla s'accepten noves construccions que substitueixen C_4 per una conjunció i X_5 per una paraula aïllada, com ara *serum and egg albumin.* El candidat a terme no és la nova construcció sencera, sinó el terme *coordinat*[22], en aquest cas: *egg albumin*. Les paraules que han donat lloc a la nova regla (*egg* i *albumin*) mantenen la seva funció d'equacions de restricció de la regla original i, a més a més, serveixen d'anclatge per a l'aplicació de la metaregla.
Una metaregla pot tenir associades restriccions específiques per limitar la seva aplicació. Per exemple:

<C_4 lema> ! but
<X_5 cat> ! Dd
<X_5 cat> ! Di

D'aquesta manera, es rebutgen les seqüències sense cap relació lèxica com és el cas de *serum and the albumin*[23].

Aquesta regla pertany a la classe de les regles de coordinació, però n'existeixen també d'inserció i de permutació.

Un exemple de regla d'inserció seria el següent:

$X_2\ X_3 \rightarrow X_2\ \mathbf{X_4}\ X_3$

[22] L'autor utilitza aquesta denominació per referir-se a una unitat polilexemàtica formada per juxtaposició.
[23] La restricció que s'aplica en aquest cas és que l'element X_5 no pot ser un determinant.

donat *medullary carcinoma*, si troba *medullary thyroid carcinoma* el terme incorporat a la llista és *thyroid carcinoma.*

I una regla de permutació:

$X_2\ X_3 \rightarrow X_3\ \mathbf{P_4}\ \mathbf{X_5}\ X_3$

donat *control center*, si troba *center for desease control* el nou terme proposat és *disease control.*

La metagramàtica de FASTR per a l'anglès inclou 73 metaregles :

- 25 de coordinació
- 17 d'inserció
- 31 de permutació

Cada metaregla està lligada a un extractor de patrons, eina que permet fer l'adquisició d'informació molt ràpida. El formalisme gramatical utilitzat per FASTR és una extensió de PATR-II (Shieber, 1986), llenguatge que permet escriure gramàtiques utilitzant estructures de trets. Les regles que descriuen els termes estan formades per una part lliure de context ($N_1 \rightarrow N_2\ N_3$) i un conjunt d'equacions que indiquen les restriccions (ex. <N2 lema>=serum i <N3 lema>=albumin). El sistema, en primer lloc, filtra les regles que ha d'aplicar en funció del text d'entrada i, a continuació, en fa una anàlisi (aquest mecanisme, però, no sempre té èxit, ja que l'èxit depèn de la paraula afegida).

Pot donar-se el cas que la metaregla permeti redescobrir un terme ja reconegut prèviament i d'aquesta manera s'estableixi un lligam entre aquestes dues unitats. Els casos d'elisions (com ara *Kerr magnetooptical effect* i *Kerr effect*) no es tacten, ja que l'aproximació escollida no és apropiada per aquest tipus de referències.

És important notar que el procés és incremental: a partir d'un conjunt de termes ja coneguts el sistema en detecta de nous, la qual cosa permet reiniciar una altra vegada el cicle i detectar més candidats. El procés continua fins que ja no es detecten nous termes.

L'autor presenta un experiment que va realitzar a partir d'un corpus de medicina d'1,5 milions de paraules i una llista de referència de 70 mil termes de diferents àrees temàtiques. Després de quinze cicles es van detectar 17 mil termes dels quals 5 mil eren nous. La velocitat de processament va ser de 2.562 paraules/minut.

Aquest sistema, però, es degrada quan la llista de referència disminueix; per exemple, si la subllista de referència de medicina es redueix a sis mil termes, es reconeixen només 3.800 nous termes.

Es planteja l'existència d'una relació conceptual entre els nous termes i el terme que ha permès el seu reconeixement. Aquesta relació varia en funció del tipus de regla aplicada: inserció o coordinació. La permutació no permet d'establir cap relació donada la naturalesa sintagmàtica de la relació.

L'autor parteix del supòsit que quan dos termes apareixen coordinats comparteixen sempre un mateix esquema semàntic. Per exemple, *dorsal spine* and *cervical spine* poden relacionar-se per coordinació i, en canvi, cap del dos apareixerà coordinat amb *fish spine* (eriçó) perquè pertanyen a classes semàntiques diferents. Per tant, es defineix una classe conceptual de la manera següent:

> *"dos termes t i t' estan en la mateixa classe si i només si existeix una cadena de coordinació entre t i t'"*

Amb aquest tipus de relació l'autor planteja la possibilitat de construir grafs que agrupin a les distintes "classes" de paraules indicant quin és el terme origen i quin/s el/s derivat/s. Per exemple la relació:

normal control → *uraemic control*

ens indica que el segon terme s'ha obtingut a partir d'una coordinació amb el primer (ha trobat la seqüència *normal and uraemic control* després de trobar -o tenir com a referència- *normal control*).

Quan es troba un nou terme per inserció, també existeix una relació conceptual entre els dos termes, a més, en aquest cas l'autor observa una gran semblança amb la taxonomia corresponent. Per posar un exemple:

	→ *superficial tumor*
malig/benign tumor	→ *mixed tumor*
	→ *mediastinal tumor*
	→ *<part-of-body> tumor*

Valoració

Els termes que s'afegeixen a la llista de termes vàlids després de cada cicle no reben cap tipus de validació. Consegüentment, aquest tractament permet afegir a aquesta llista termes no vàlids que poden donar lloc, en els cicles successius, a incorporar més termes no vàlids. L'autor opina[24] que aquest fet no és una font d'error important, el sistema en certa manera s'autocorregeix ja que "normalment" candidats incorrectes no produeixen nous possibles candidats a termes.

És veritat, però, que el fet de poder utilitzar els termes ja reconeguts i acceptats és molt útil encara que, com reconeix l'autor, limita en certa manera la possibilitat de reconèixer termes que no estan relacionats amb els de l'origen. Segurament que aquesta tècnica no hauria de tenir valor únic, sinó que hauria d'ésser complementària d'altres estratègies.

[24]Comunicació personal.

2.9 Heid

Autors**:** Heid, U. et al.
Publicació**:** *Term extraction with standard tools for corpus exploration*
Data**:** 1996
Objectiu**:** Demostrar la utilitat de les eines estàndards per a l'exploració lingüística dels corpus textuals per permetre l'extracció de candidats a termes
Llengües de treball**:** Alemany (anglès i francès)

Sinopsi

Heid i al. consideren que l'extracció automàtica de termes té diverses aplicacions, i que l'elaboració de diccionaris o vocabularis en seria una de les principals. Precisament, en la construcció de diccionaris a partir de corpus informatitzats, hi intervenen dos tipus de fases:

1. el preprocessament lingüístic
2. l'anàlisi de tasques específiques

Cada una d'aquestes fases requereix unes eines informàtiques específiques.

1. Preprocessament lingüístic

En la fase de processament lingüístic es serveixen de les eines següents:

1. un segmentador o *tokening*, que identifica els límits de les frases i dels mots.
2. un analitzador morfosintàctic, que identifica les categories gramaticals i les característiques distribucionals i morfosintàctiques.
3. un etiquetador o *tagging*, que desambigua les hipòtesis morfosintàctiques
4. un lematitzador, que identifica el lema dels candidats

Els autors destaquen que l'analitzador morfosintàctic per a alemany encara no està acabat del tot.

2. La identificació de termes

El sistema que utilitzen per a l'extracció de termes no està exclusivament dissenyat per a la terminologia, sinó que utilitzen eines informàtiques com:

1. un processador d'interrogació de corpus general (CQP)
2. un macroprocessador de CQP
3. XKWIC, per extreure concordances i extreure llistes de freqüència absoluta o relativa d'ítems buscats.

L'extracció de termes està lligada a un llenguatge d'interrogacions complex. Aquestes interrogacions seran diferents en funció del tipus de terme que es vulgui buscar. Així, per exemple, les interrogacions sobre termes simples es fan a partir dels morfemes o dels components típics dels

compostos o derivats. Parteixen del supòsit que els termes nominals afixats dels llenguatges d'especialitat utilitzen uns determinats sufixos o/i prefixos més que uns altres. Pel que fa a l'extracció de seqüències de paraules (N-V), en canvi, les interrogacions es basen en patrons.

Aquests autors han aplicat aquest conjunt d'eines a una mostra de 35.000 mots de textos tècnics sobre enginyeria de l'automòbil en alemany. Els resultats són els següents:

- Quant als termes simples: 90% de candidats a terme i un 10% de silenci. Dintre dels termes candidats els autors han extret la quantitat de soroll per a cada tipus d'esquema i han comprovat que aquest percentatge varia considerablement d'un esquema a un altre.
- Per als termes complexos no donen resultats concrets, sinó que es limiten a dir que són més insatisfactoris i que responen als mateixos problemes de les gramàtiques regulars. Per exemple, creuen que si es tingués un analitzador sintàctic, com posseeix l'anglès, el soroll disminuiria.
- Finalment, en l'extracció de col·locacions[25] es demostra que es produeix soroll, però no silenci, pel fet que es basen en el criteri de la freqüència.

Els autors han aplicat al mateix corpus de 35.000 mots el programa estadístic d'Ahmad i han comprovat que els resultats extrets amb un interrogació lingüística estan inclosos en l'output fruit dels mètodes estadístics, però el soroll és molt superior en els mètodes estadístics que en els lingüístics.

Valoració

Destaquem les observacions següents:

Per atendre el sistema que fan servir s'han de tenir presents les característiques morfosintàctiques de la llengua alemanya. L'alemany, en contraposició amb altres llengües com ara les romàniques, tendeix a formar els seus compostos de manera sintètica. És a dir, allò que en altres llengües s'expressa mitjançant sintagmes terminològics, en alemany s'expressa amb una sola paraula, entenent per paraula tot segment entre dos blancs. Per això, en alemany la dificultat de l'extracció automàtica de termes no radica tant en la delimitació dels termes, com en la naturalesa terminològica d'un mot i, en conseqüència, necessiten paràmetres per distingir un terme d'un mot de la llengua general que té la mateixa estructura morfosintàctica interna.

Com la majoria de programes examinats, se centra en els termes nominals, encara que aquest sistema també pot extreure col·locacions formades per la combinació d'un nom i un verb. En aquest darrer cas, els resultats són, segons els autors, molt pitjors. No tenim dades exactes del funcionament i dels resultats del sistema.

[25]En aquest estudi utilitzarem el terme *col·locació* per referir-nos a les concurrències fraseològiques.

2.10 LEXTER[26]

Autor**:** Bourigault, D.
Publicació**:** *Lexter, a Terminology Extraction Software. Application to Knowledge Acquisition from texts*. PhD Thesis, École des Hautes Études en Sciences, París
Data**:** 1994
Objectiu**:** Eina d'extracció de candidats a terme
Llengües de treball**:** Francès (i anglès)

Sinopsi

L'origen d'aquest sistema s'ha de buscar en les necessitats de la societat EDF (*Electricité de France*) de millorar un sistema d'indexació de textos ja existent. La idea bàsica és la detecció de les fronteres entre les quals s'espera aïllar els sintagmes nominals susceptibles de ser considerats unitats terminològiques. L'anàlisi que realitza LEXTER és superficial i fa servir heurístiques pròpies del text a tractar per tal d'obtenir sintagmes nominals de longitud màxima que considera candidats a terme.

Encara que el funcionament d'aquest sistema es basa exclusivament en tècniques lingüístiques obté uns resultats considerablement bons.

> *LEXTER takes advantage of negative knowledge about the form of complex terms, by identifying those surface patterns which never go to make up these terms and which can thus be considered as potential terminological limits*

El programa està organitzat al voltant de cinc mòduls principals:

- mòdul d'anàlisi morfològica i desambiguació
- mòdul de delimitació
- mòdul de descomposició
- mòdul d'estructuració
- mòdul de navegació

1. En el **mòdul d'anàlisi morfològica i desambiguació,** els textos per tractar reben informació de la categoria gramatical i del lema de cada una de les paraules. Aquestes són les úniques informacions externes que rep el sistema.

2. En el **mòdul de delimitació**, s'elabora una anàlisi sintàctica local per descompondre el text en grups nominals de longitud màxima. Per exemple:

> *alimentation en eau*
> *pompe d'extraction*
> *alimentation electrique de la pompe de refoulement*

[26] Logiciel d'EXtraction de TERminologie

En aquest moment el sistema s'aprofita del coneixement negatiu sobre la composició d'un terme complex. Amb aquest objectiu identifica el patrons que mai formaran part d'un terme, com per exemple els verbs en forma personal, pronoms, conjuncions, etc. i els considera com una frontera del possible terme. Alguns d'aquests patrons són simples (pronom, verb en forma personal, etc.), mentre que altres són patrons complexos (seqüències de preposició + determinant). Un exemple en francès d'aquest últim cas és la seqüència: 'SUR(prep) + LE(article definit)'.

L'anàlisi més habitual és considerar que aquesta seqüència marca una frontera entre grups nominals, com per exemple a:

on raccorde le câble d'alimentation du banc sur le coffret de décharge batterie

Però existeix un número de casos (10%) en què aquesta seqüència forma part del terme, com ara:

action sur le bouton poussair de réarmement
action sur le système d'alimentation de secours

Per resoldre aquesta i altres situacions semblants[27], el sistema fa servir un mètode d'aprenentatge endogen dels patrons de subcategorització. Aquesta estratègia consisteix a recórrer el corpus per trobar les seqüències *(nom)+Sur+le* amb el context a la dreta. A continuació s'eliminen els noms que no són productius (que no tenen contextos a la dreta diferents). En la segona passada, les seqüències *sur+le* són considerades fronteres de frase excepte en els casos que van precedides pels noms productius detectats en la fase d'aprenentatge.

Per exemplificar aquest funcionament suposem que en una primera passada el sistema es troba amb seqüències com les següents:

*le **protection contre** le gel est assurée par*
***Protection contre** les grands froids*
*il s'agit de maintenir la teneur en oxygène de cette **eau dans** les limites fixées*
*on procède à l'injection d'**eau dans** les générateurs de vapeur*
*le système permet l'aiguillage des **automates sur** le prélèvement effectué*

La productivitat de cada seqüència *Sur* és

protection contre	2
eau dans	2
automates sur	1

En la segona passada aquelles seqüències productives (freqüència més gran que 1) no són considerades pel sistema com una frontera de terme, mentre que les seqüències no productives sí

[27] En aquest grup l'autor situa las preposicions següents: *avec*, *contre*, *dans*, *par*, *pour*, *sans*, *sous*, *vers*, *en* i *sur*. Així, es parla d'una classe de preposicions que anomena *Sur* i no de la preposició *sur* en concret.

que són considerades com un límit extern del candidat a terme. En l'exemple que hem presentat *protection contre* i *eau dans* no actuarà com a frontera mentre que *automates sur* si que ho serà.

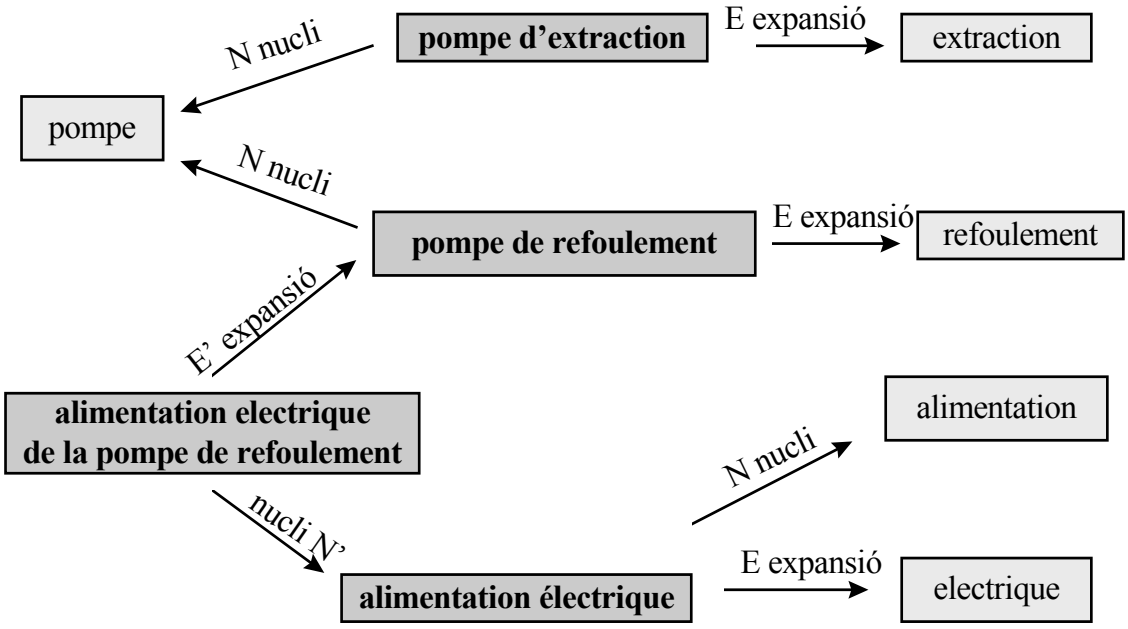
Així, amb aquesta estratègia es detecta positivament un nombre important de noms que d'altra manera s'haurien perdut. Però amb aquesta permeabilitat s'introdueix una quantitat de brutícia considerable, estimada per l'autor entre el 10% i el 50 %.

3. En el **mòdul de descomposició,** s'analitzen els grups nominals i se'n divideixen els constituents en nucli i expansió:

pompe d'extraction →	*nucli:*	*pompe*
	expansió:	*extraction*

En aquest punt el sistema es troba amb situacions d'ambigüitat com ara les seqüències: "Nom Adj_1 Adj_2" o "Nom_1 Prep Nom_2 Adj" que no se sap com s'ha d'analitzar. Per a la resolució d'aquests casos es recorre a un sistema d'aprenentatge endogen similar al que hem presentat en el mòdul de delimitació.

4. En el **mòdul d'estructuració,** s'organitza la llista de candidats a terme en una xarxa anomenada xarxa terminològica. Per a l'obtenció d'aquesta xarxa és suficient recórrer a la llista de candidats a terme i reconèixer les distintes parts de cada candidat a terme. Presentem a continuació un exemple d'aquestes xarxes:



Aprofundint en l'anàlisi dels candidats a terme l'autor desglossa els tipus d'enllaços existents entre les diferents parts d'un candidat a terme. Així, en el corpus DSE LEXTER troba 16526 enllaços (productivitat global p) que es reparteixen de la manera següent:

5463 tipus N (pN)
5463 tipus E (pE)
2800 tipus N' (pN')
2800 tipus E' (pE')

Aquesta classificació permet una agrupació dels components dels candidats a terme en funció de la posició que ocupen (N, E, N' i E'). Amb aquestes dades també es defineix la *taxa de productivitat normal* i la *taxa de productivitat ponderada*:

taxa de productivitat	
normal	ponderada
xN = pN/p	cN = xT log(pN)
xE = pE/p	cE = xE log(pE)
xN' = pN'/p	cN' = xT' log(pT')
xE' = pE'/p	cE' = xE' log(pE')

Aquests coeficients no s'utilitzen com a filtre, sinó que es presenten al terminòleg com una dada més per facilitar l'avaluació dels candidats a terme.

5. Finalment, en el **mòdul de navegació**, es construeix una interfície de consulta anomenada *hipertext terminològic* a partir del corpus inicial, de la xarxa de candidats a termes i dels coeficients i llistes ja mencionades.

LEXTER s'utilitza actualment en l'explotació de diferents corpus de la societat EDF, bàsicament en la indexació automàtica de textos, en els sistemes de consulta hipertextuals de documentació tècnica, en l'adquisició de coneixement i en la construcció de bases de dades terminològiques.

Actualment, LEXTER també s'utilitza com a extractor de terminologia en la Base de coneixements terminològics dissenyada pel grup de terminologia de TIA i a SYCLADE (Habert, 1996), una eina per a la classificació de paraules. En aquest darrer cas, a partir d'un candidat a terme, el sistema simplifica l'arbre d'anàlisi complex en arbres elementals. Així, de l'arbre d'anàlisi del candidat a terme:

sténose serrée de le (du) tronc commun gauche

s'obtenen els arbres elementals de les seqüències següents:

- *sténose serrée*
- *sténose de tronc*
- *tronc commun*
- *tronc gauche*

Aquests arbres elementals permeten establir classes de contextos sintàctics. Del primer arbre elemental (*sténose serrée*) es poden establir dos contextos possibles: ~ *serrée* i *sténose* ~, on "~" representa el nucli del sintagma. Les paraules nucli es convertiran en el nusos d'un graf mentre que el context servirà d'etiqueta als costats. Per exemple, a partir dels candidats a terme *sténose sévère* i *lesion sévère*, es pot establir la relació següent:

esténose —sévère— lesion

La idea subjacent és que els costats relacionen paraules semànticament pròximes. Si cada costat té associada la freqüència d'aparició de cada arbre elemental, la profunditat del graf variarà en funció del sostre escollit. Això fa surar les paraules més pròximes com: *malalties, actes mèdics*, *parts del cos*, *grau d'afecció*, *relacions "part de"*.

Valoració

La capacitat del sistema es ressent greument d'errors en el marcatge i en la desambiguació prèvia. C. Jacquemin apunta que aquest sistema (com tots els que utilitzen tècniques simbòliques) genera una quantitat notable de soroll: d'un corpus de 200.000 paraules s'obtenen 20.000 candidats a termes que després de la validació es redueixen a 10.000. L'autor[28] remarca també el problema del silenci (termes no reconeguts) que valora com el 5% del total dels termes vàlids.

Com la majoria de sistemes, es limita a la detecció de sintagmes nominals, ja que els verbs són considerats frontera de terme i mai s'incorporen.

Un dels punts més valorats del sistema és el mecanisme d'aprenentatge endogen que permet treballar autònomament sense necessitat de tenir accés a un diccionari complex i voluminós. També hem de destacar la utilitat de la presentació dels resultats en format hipertextual ja que facilita enormement la tasca del "terminòleg".

2.11 NEURAL

Autors**:** Frantzi, K. T.; Ananiadou, S.
Publicació**:** *Statistical measures for terminological extraction*
Data**:** 1995
Objectiu**:** Eina d'extracció de candidats a terme
Llengua de treball**:** Anglès

Sinopsi

[28] Bourigault, D. (desembre, 1996), *Detecció de Termes : estat de la qüestió*. Jornada de treball, Institut universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona

Neural és un sistema de naturalesa híbrida d'extracció d'unitats terminològiques en què es combinen dues mesures estadístiques (informació mútua i freqüència) amb coneixement lingüístic (regles morfològiques) mitjançant xarxes neuronals.

És sabut que l'ús de la freqüència com a única mesura de la fiabilitat d'un candidat a terme tendeix a afavorir només l'extracció dels termes més freqüents d'un document i, en canvi, no detecta els més esporàdics. Per superar aquesta dificultat l'autor utilitza també la freqüència mútua, amb una fórmula modificada per tenir en compte els termes de longitud igual o major a dos. El càlcul de la freqüència s'hauria de modificar si es volen tenir en compte les *concurrències niades,* per exemple en seqüències candidates a terme com les següents:

a) *hay fever conjunctivitis*
b) *fever conjunctivitis*
c) *hay fever*

Si el sistema extreu la primera opció *hay fever conjunctivitis* com a candidat a terme, també ha d'extreure la segona i la tercera opció ja que $f(b) \geq f(a)$ i $f(c) \geq f(a)$. El problema radica que b) no és un terme, sinó que només és una col·locació.

En contrast tenim seqüències com ara:

- *disposable soft contact lenses*
- *cosmetic soft contact lenses*
- *therapeutic soft contact lenses*
- *standard hard contact lenses*
- *gas permeable hard contact lenses*

En totes aquestes seqüències apareix el segment *contact lenses,* cosa que facilita la detecció. Els autors afirmen que el problema és molt complex i no està resolt i apunten també que només té sentit estudiar el problema de la detecció de termes en textos especialitzats. En aquests, a diferència dels textos de tipus general, no es presenta el problema de la insuficiència de dades, problema que es presenta sobretot quan es treballa amb mètodes estadístics.

El corpus de treball d'aquest estudi està constituït per 55.000 paraules de l'àrea de medicina, concretament de la branca d'oftalmologia.

Els autors de NEURAL es limiten a analitzar dues estructures:

- Nom-Nom
- Adjectiu-Nom

El sistema disposa d'una llista de sufixos que es donen amb molta freqüència en les unitats terminològiques del domini de l'oftalmologia com ara *-oid, -oma, -ium*.

En les proves realitzades sobre textos molt especialitzats[29] d'oftalmologia s'ha obtingut un 70% d'encert. El sistema utilitza una xarxa neuronal del tipus *back propagation* de dos capes. Actualment, els autors treballen en un projecte per augmentar les dimensions del corpus de 55.000 a 500.000 paraules i optimitzar el funcionament de la xarxa neuronal.

2.12 NODALIDA-95

Autor**:** Arppe, A.
Publicació**:** *Term extraction from unrestricted text*
Data**:** 1995
Objectiu**:** Eina d'extracció de candidats a terme
Llengua de treball**:** Anglès
Sinopsi

Nodalida és un producte dissenyat per l'empresa Lingsoft a partir de l'eina NPtool[30]desenvolupada a la Universitat de Hèlsinki, Departament de Lingüística General. NPtool genera llistes dels sintagmes nominals de les frases d'un text (NP) i proporciona una avaluació de la certesa d'aquests sintagmes com a candidats a termes (ok/?). D'aquestes llistes s'extreuen totes les subcadenes permissibles i es multiplica per tres la llista inicial. Així, per exemple, per a la frase:

exact form of the correct theory of quantum gravity

NPtool proposa la següent llista de NPs addicionals:

exact form of the correct theory
form of the correct theory of quantum gravity
form of the correct theory
theory
gravity
exact form
form
correct theory
quantum gravity

Paral·lelament operen una sèrie de premisses, com la següent, que serveixen com a primer filtre:

> "Els NPs que comencin amb determinant, adjectiu o una frase prefixada (*kind of*, *some*, *one*,...) són eliminats"

Per a la resta de NPs, se'n calcula la freqüència d'aparició i s'ordenen i agrupen segons el nucli gramatical i la freqüència i es presenten al terminòleg acompanyats del context.

[29] Els autors pretenen d'aquesta manera resoldre el problema de la manca d'informació (*sparse data*), associat amb tots els sistemes que utilitzen tècniques estadístiques.
[30] Aquesta eina s'utilitza també en el marc del projecte TRANSTERM, finançat per la Unió Europea. Per més informació podeu consultar Ahmad, 1996.

El mètode de presentació és molt semblant a l'utilitzat a *Termight*. D'aquesta manera, el terminòleg pot determinar quin/s del NP proposats és susceptible de ser considerat terme.

L'eina NPtool (Voutilanen, 1993), el cor del sistema, és un detector de sintagmes nominals basat en gran mesura en el formalisme de *gramàtica de restriccions* (*constraint grammar*) (Karlsson, 1990). Té les següents característiques bàsiques:

- L'anàlisi morfològica es basa en una descripció molt acurada que inclou categoria morfològica i funció sintàctica.
- La descripció morfològica es basa en regles lingüístiques. S'eviten, però, distincions que depenguin del coneixement a un nivell superior del sintàctic.
- Es considera que, tant en la gramàtica com en el lexicó, l'anàlisi s'ha de poder aplicar a un corpus amb text no controlat.
- La desambiguació s'efectua amb criteris estrictament lingüístics. Aquest procés deixa ambigües entre un 3% i 6% del total de paraules.

El text és sotmès a un procés previ per determinar les fronteres de frase, locucions, compostos, signes tipogràfics, etc. A continuació, s'analitza morfològicament i s'obté un resultat com el següent[31]:

*("<*the>"*	*("the" DET CENTRAL ART SG/PL (@>N)))*
("<inlet>"	*("inlet" N NOM SG))*
("<and>"	*("and" CC (@CC)))*
("<exhaust>"	*("exhaust" <SVO> V SUBJUNCTIVE VFIN (@V))*
	("exhaust" <SVO> V IMP VFIN (@V))
	("exhaust" <SVO> V INF)
	("exhaust" <SVO> V PRES -SG3 VFIN (@V))
	("exhaust" N NOM SG))
("<manifold>"	*("manifold" N NOM PL))*

En aquest moment, es produeix la desambiguació, per exemple per a la frase:

The inlet and exhaust manifolds are mounted on opposite sides of the cylinder head.

s'obtenen les dues anàlisis següents:

a) ... on/@AH opposite/@N sides/@NH of/@N< the/@>N ***cylinder/@NH head/@V***
b) ... on/@AH opposite/@N sides/@NH of/@N< the/@>N ***cylinder/@>N head/@NH***

[31]El significat dels símbols utilitzats és el següent: @>N = premodificadors, @<N postmodificadors, @ conjuncions de coordinació i subordinació, V verb i marcador d'infinitiu, NH nucli nominal, "<" i ">" indiquen la direcció en què es troba el nucli, esquerra o dreta.

El fet de considerar o no la seqüència final (*cylinder head*) com un sintagma nominal constitueix l'única diferència entre les dues anàlisis proposades. El procés que segueix només dóna dues anàlisis possibles per a cada frase: una en què es dóna preferència als NPs de longitud màxima (*NP-friendly*) i una altra en què es dóna preferència als NPs de longitud mínima (*NP-hostile*). A continuació, el sistema efectua una comparació de les dues estratègies i assigna una valoració d'aquest tipus :

ok : la mateixa anàlisi és proposada per les dues estratègies
? : l'anàlisi és proposada només per a una de les estratègies

Així , per a l'última frase el sistema dóna les anàlisis següents:

ok : inlet and exhaust manifolds
ok : exhaust manifolds
? : opposite sides of the cylinder
? : opposite sides of the cylinder head

El terminòleg rep una llista de candidats a terme per validar amb aquesta informació addicional. Els resultats obtinguts per l'eina auxiliar NPtool son força bons (precisió = 95-98%, recall = 98,5-100%) amb un text de 20.000 paraules aprox.

Valoració

El sistema NODALIDA es basa en l'ús de coneixement lingüístic mitjançant una aproximació estructural (detecció de estructures sintagmàtiques i corresponent desambiguació estructural).
Els resultats presentats per l'autor impliquen un alt grau de qualitat encara que s'haurien d'augmentar les dimensions del corpus perquè fins ara totes les proves s'han realitzat sobre corpus força reduïts. Tampoc queda clar com l'autor calcula las xifres de precisió i recall, en particular quin són els termes considerats correctes: els que tenen el signe ok o tots. S'ha de fer notar també que aquests resultats es refereixen a l'eina auxiliar NPtool i no al sistema NODALIDA en situacions reals d'extracció de terminologia.
Tenint present que el desambiguador és una de les principal fonts d'error en aquests sistemes, els autors creuen haver obtingut un alt grau de qualitat, encara que no existeixen dades en una aplicació d'extracció de termes real . A més, per aconseguir aquesta qualitat el sistema fa ús d'un número molt elevat de regles, que comporten un problema considerable de gestió i de control.
La llista que el sistema proposa al terminòleg per validar són tots els candidats valorats amb els signes **ok** o **?**. La manera d'obtenir tots els possibles sintagmes nominals que utilitza aquest sistema fa pensar que en la llista a validar hi ha molts candidats a terme que el terminòleg ha de rebutjar.

2.13 SBIC

Autor**:** Anzaldi, C.
Publicació**:** *Prototipo di thesaurus per l'energia e l'ambiente tramite il sistema SBIC*
Data**:** 1996

Objectiu**:** Indexació de documents
Llengua de treball**:** Italià
Sinopsi

SBIC ha estat desenvolupat bàsicament per servir com a thesaurus d'un sistema de gestió documental. En opinió de la seva autora, però, pot ser utilitzat també per a la detecció de terminologia.

La seva característica principal és la de treballar amb un corpus textual sense fer ús de pràcticament cap informació lingüística. Aquest fet comporta que el processament estrictament informàtic sigui molt ràpid, contràriament deixa per al terminòleg/especialista una tasca important de selecció i composició manual dels termes pertinents.

La idea de base és semblant a la del sistema LEXTER: s'estableix a priori una llista de paraules que serviran de frontera als termes potencials, però, com que no disposa d'informació lingüística, extreu fragments de text assimilables a sintagmes nominals i després el terminòleg haurà de seleccionar manualment els que consideri vàlids.

Valoració

L'ús escàs que fa aquest sistema de coneixement lingüístic i el fet que el terminòleg ha d'intervenir en diverses etapes fa difícil considerar-lo profitós per a l'extracció de terminologia.

2.14 Termight

Autors**:** Dagan I., Church K.
Publicació**:** *Termight: Identifying and Translating Technical Terminology*
Data**:** 1994
Objectiu**:** Eina d'extracció de candidats a terme
Llengües de treball**:** Anglès i alemany

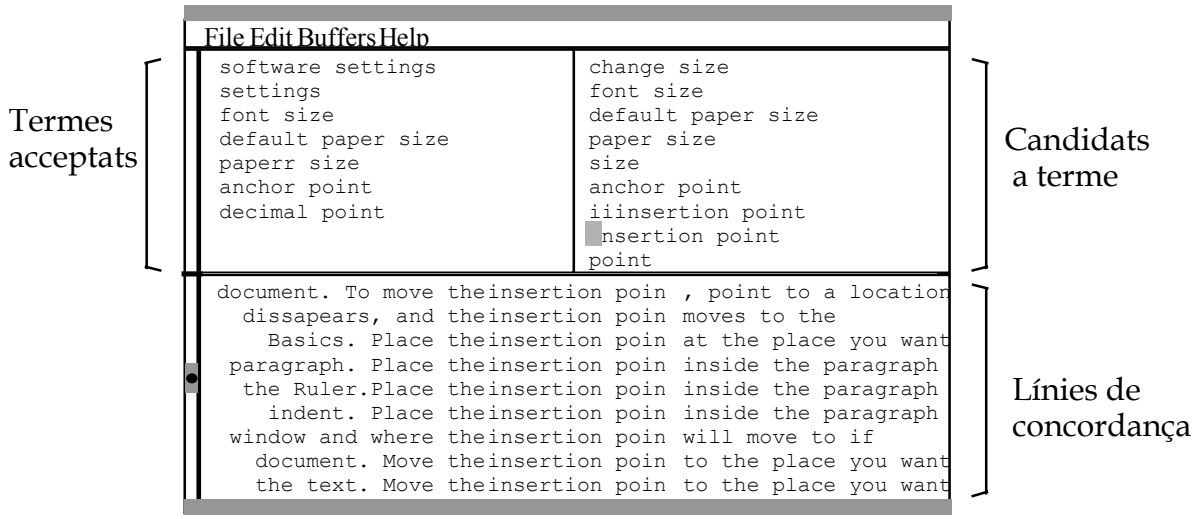
Sinopsi

Termight és un sistema que van desenvolupar I. Dagan i K. Church i que actualment s'utilitza en les oficines d'*AT&T Business Translation Systems*. Ha estat concebut com una eina per automatitzar algunes etapes de la recerca terminològica que efectuen els traductors professionals.

Per portar a terme aquesta tasca, es parteix d'un text ja etiquetat i desambiguat i d'una llista de patrons sintàctics predefinits que es poden ajustar per a cada document que es vol explorar. D'aquesta manera el sistema obté una llista de candidats a terme formada per seqüències d'una o més paraules. Els candidats formats per una única paraula es defineixen com totes les paraules que no estan incloses en una llista de paraules buides o *stop list* definida prèviament. D'altra banda, els termes compostos per dues o més paraules es corresponen amb un dels patrons sintàctics predefinits

mitjançant expressions regulars; en les seves proves els autors van limitar aquests patrons a seqüències de noms.

El candidats a terme obtinguts s'agrupen i classifiquen en funció del seu lema (el nom situat a la dreta) i la seva freqüència. Els candidats amb un mateix lema es classifiquen alfabèticament en funció de l'ordre invers de les paraules que el composen, d'aquesta manera, es reflecteix l'ordre de modificacions de les frases nominals simples en anglès.



Per a cada candidat a terme s'obtenen les concordances corresponents classificades alfabèticament en funció del seu context. Amb aquesta informació es crea una pantalla per permetre que el terminòleg avaluï fàcilment la pertinença o no de cada candidat.

Els autors expliquen que s'obté un rendiment d'extracció entre 150 i 200 termes/hora, xifra que dobla el ritme habitual manual. Des del punt de vista de la qualitat d'extracció, els autors afirmen que aquest mètode, a diferència dels sistemes exclusivament estadístics, permet extreure també els termes d'ocurrència molt baixa.

A més, el sistema disposa d'un mòdul bilingüe. Aquest component obté, per mètodes estadístics, una alineació a nivell de paraula dels textos. A partir d'aquesta paral·lelització, es fan correspondre els termes trobats en una llengua A amb l'equivalent en una llengua B. Aquesta llista de candidats a terme, oportunament classificada, es presenta una altra vegada al terminòleg per a la seva avaluació.

Aquest segon mòdul de Termight, però, no ha estat tan desenvolupat i provat com el primer. Les proves s'han realitzat amb 192 termes d'un manual tècnic en anglès i en alemany. La traducció correcta es troba el 40% de les vegades en primer lloc i en el 7% del casos en segon lloc de les possibilitats suggerides. En la resta de casos, la traducció correcta estava en algun dels altres llocs de la llista de propostes.

Valoració

Destaquem les observacions següents:

Des del punt de vista dels aspectes a millorar s'ha de remarcar que:

- No es dóna cap informació numèrica sobre la qualitat del reconeixement o del silenci.
- L'únic patró sintàctic previst és molt simple: *seqüències de noms*. Aquest patró pot ser més o menys vàlid per a l'anglès, però per a qualsevol llengua romànica s'hauria de complicar molt més i arribaríem, segurament, als patrons previstos en altres propostes. La terminologia detectada, per tant, es limita als sintagmes nominals.
- L'aplicació als textos bilingües, tot i ser molt interessant, no obté bons resultats probablement per la dificultat intrínseca de la paral·lelització de textos a nivell de paraula.

En contraposició, s'ha de valorar com a suggerents els aspectes següents:

- La classificació molt acurada dels candidats a termes.
- El fet de no pretendre ser un sistema totalment automàtic, sinó una ajuda al traductor.
- El sistema d'ordenació i presentació dels candidats a terme amb els seus contextos corresponents en finestres per poder-los seleccionar sembla molt encertat.

2.15 TERMINO

Autors**:**	Plante, P. et al. (Centre ATO)
Publicació**:**	Diversos articles (vegeu bibliografia)
Data:	1990 (primera versió)
Objectiu**:**	Facilitar les tasques del terminògraf: delimitació de termes, extracció, creació de fitxes terminològiques, creació de bases de fitxes terminològiques. Es pretén crear un lloc de treball integrat per al terminòleg.
Llengua de treball**:**	Francès

Sinopsi

TERMINO és un programa que integra un conjunt d'eines per facilitar l'extracció de termes de textos francesos. Representa una ajuda per al terminòleg en la identificació de les unitats de discurs que designen nocions o objectes i dóna per a cada una d'aquestes unitats el context immediat del qual es poden extreure dades relatives a les nocions designades per aquestes unitats.

Existeixen diverses versions de TERMINO que milloren algun aspecte o incorporen alguna característica nova respecte de l'anterior.

El tractament d'aquesta eina recolza bàsicament en coneixements lingüístics i està format per tres subsistemes:

- un preeditor, que separa el text en paraules i frases i marca els noms propis
- un analitzador morfosintàctic
- un facilitador de *record-draftings*

El text no ha de rebre cap tractament especial, només ha d'estar codificat en ASCII.

Per a la delimitació i extracció de termes el subsistema més interessant és, però, l'analitzador morfosintàctic.

Subsistema 2: analitzador morfosintàctic

L'analitzador morfosintàctic està compost de tres mòduls:

1. analitzador morfològic
2. analitzador sintàctic o *parser*
3. detector de sinapsis

1. L'analitzador morfològic cobreix dues funcions:

- la categorització automàtica
- la identificació del lema i l'etiquetador

Segons els autors el 30%[32] del mots en francès poden pertànyer a més d'una categoria, en conseqüència els autors han decidit adoptar una estratègia àmplia i etiquetar totes les possibilitats categorials per a cada mot, per això durant aquesta fase es sobreprodueix mots amb etiquetes diferents.

La categorització i la lematització s'obtenen amb l'aplicació del programa LCMF (*Lemmatisation et catégorisation morphologique du français*) creat per Lucie Dumas i Pierre Plante. Segons els autors, LCMF no és un diccionari, sinó que es tracta d'un analitzador morfològic de formes lèxiques i, per tant, pot lematitzar i etiquetar correctament una forma lèxica nova.

2. L'analitzador sintàctic (*parser*)

L'analitzador sintàctic s'encarrega d'eliminar la majoria d'ambigüitats que s'havien generat en la darrera fase mitjançant la construcció d'una estructura sintàctica per a cada frase. L'analitzador sintàctic utilitzat és l'ALSF (A*nalyseur lexico-syntaxique du français*) creat per Jean-Marie Marandin, Sophie David i Pierre Plante. Està escrit en llenguatge FX, llenguatge per a la programació *en faisceaux* desenvolupat per Pierre Plante.
3. Detector de sinapsis

[32]Aquesta xifra, que els autors no especifiquen a partir de quin tipus de font ha estat calculada, resulta sorprenent si la comparem en l'índex d'ambigüitat del català o del castellà (entre el 60% i el 70%).

El detector de sinapsis (MRSF) selecciona, entre les unitats sintàctiques que surten del *parser*, les unitats nominals lèxiques susceptibles de ser termes.

El MRSF és un mòdul creat específicament per a TERMINO per Sophie David que està subdividit en cinc mòduls:

1. un mòdul de nucli
2. un mòdul de construcció d'expansions
3. un mòdul de categorització
4. un mòdul de generació
5. un mòdul de representació i avaluació

El MRSF recolza en principis de construcció del grup nominal. S. David entén per sinapsi[33] una unitat polilèxica d'origen sintàctic que constitueix el nucli d'una frase nominal. Les sinapsis són exclusivament grups nominals, entre les quals n'hi haurà algunes que seran termes i d'altres que no ho seran. Algunes seran només "temes" que permetran al terminòleg conèixer diferents conceptes o tenir una idea general de les perspectives temàtiques del text.

EL mòdul de nucli busca totes les sinapsis, el mòdul de construcció d'expansions analitza les expansions i cada una rep una anàlisi sintàctica. Per exemple "*El sistema de gestió de bases de dades...*" quedaria desglossat en:

sistema de gestió de bases de dades
gestió de bases de dades
bases de dades
dades

El mòdul de categorització comprèn tres tipus d'informacions:

- la pertinença o no d'una extensió a un nucli
- la posició o no de l'expansió en un nus de l'arbre sintàctic
- la presència d'una organització lineal de la sinapsi

Aquestes informacions es transmeten al mòdul de generació, que s'ocupa d'associar un nucli a les extensions. Finalment, l'últim mòdul avalua les sinapsis a través de regles d'enriquiment de representacions de sinapsis. A més, un grup de regles heurístiques serveixen per reforçar positivament o negativament la tria d'un candidat a partir de criteris de freqüència, de categorització d'expansions o de criteris lèxics. Alguns exemples d'aquest tipus de regles són les següents:

- "si el nucli és un nom propi o una sigla, no admet cap extensió"

[33]La noció de *sinapsi* en S. David és més àmplia que la de Benveniste, 1974.

- "s'han d'excloure les sinapsis el nucli de les quals sigui un lexema del tipus "terme", "mot", "absència", "presència", etc."
- "s'han d'excloure les sinapsis l'extensió de les quals sigui adjectival i contingui un mot com "diferent", etc."

El conjunt d'eines informàtiques que ofereix TERMINO és molt més ampli i comprèn diferents mòduls que permeten manipular les dades terminològiques, com per exemple el mòdul MRF de redacció assistida de fitxes (aquest, però, és semiautomàtic).

Totes aquestes eines ajuden el terminòleg a:

- decidir-se si una sinapsi és un terme o no
- confeccionar fitxes terminològiques
- crear bases de dades terminològiques

TERMINO reconeix del 70% al 74% per cent dels termes complexos. El fet que TERMINO no reconegui gairebé un 30% del termes és bàsicament per tres factors:

- la coordinació (ja que és una marca de trencament de segment)
- els acrònims
- els noms comuns amb majúscula

A més, el sistema produeix un 28% de soroll, del qual el 47% és causat per un etiquetatge incorrecte i un 53% és degut a les sinapsis que pertanyen al llenguatge general.

> *"Noise results from the arbitrary decision to output only right-branching nested synapsies, from parsing errors caused by the lack of recognition of frequently occurring idiomatic prepositional phrases and particularly from what would appear to be the inherent limitation of syntactic analysis. Indeed, terminologists are left with a considerable amount of manual filtering out of genuine synapsies without terminological status. " Lauriston (1994)*

El primer tipus de soroll segons Lauriston (1994) es podria millorar si es tinguessin en compte el adjectius amb poc poder de designació o els participis. Pel que fa al segon tipus de soroll, es necessitaria una anàlisi semàntica, treballar en camps concrets i sobre tipus de textos determinats.

Valoració

Les principals observacions del sistema TERMINO són:

- TERMINO és un dels primers extractors de candidats a termes que va funcionar, posteriorment s'han elaborat diferents versions que no hem analitzat i que de ben segur milloren l'original.
- És un extractor de base lingüística, format per diferents mòduls independents.
- TERMINO es basa en el concepte de *sinapsi* que equival a grup nominal.

- El detector de sinapsis es basa en l'establiment d'un conjunt de regles heurístiques, que segurament es podrien ampliar si es delimitava el corpus.

S'hauria de seguir perfeccionant sobretot tenint en compte que:

- El sistema encara produeix molt de silenci, cosa que podria solucionar-se amb un tractament diferent de les majúscules o dels acrònims, per exemple.
- Hi ha un 28% de soroll entre els candidats a terme.

2.16 TERMS

Autors**:** Justeson, J.; Katz, S.
Publicació**:** *Technical terminology: some linguistic properties and an algorithm for identification in text*
Data: 1995
Objectiu**:** Eina d'extracció de candidats a terme
Llengua de treball**:** Anglès

Sinopsi

Els autors de TERMS parteixen de les següents idees en relació als termes:

- Els sintagmes nominals terminològics (SNT) es diferencien dels no terminològics (SNnT) pel fet que els modificadors que formen part dels primers són molt més reduïts que els que poden aparèixer en els segons.
- Una entitat introduïda amb un SNnT pot ser referenciada més endavant només pel nucli del SN i sovint mitjançant altres SN (sinònims, hipònims, hiperònims). En contrast, una entitat introduïda amb un SNT normalment es repeteix en forma idèntica en un mateix document, ja que la sola omissió d'un modificador podria fer que l'entitat referenciada canviés.
- Segons els autors, en els textos tècnics els SN són gairebé exclusivament terminològics.
- Els termes tècnics estan formats gairebé sempre per SNs (entre el 92,5 i el 99% d'una mostra aleatòria de 800 termes són SN).
- el SNT estan formats bàsicament per noms i adjectius (97%) i algunes preposicions (3%) sempre entre dos SN.
- La longitud mitjana d'un SNT és de 1,91 paraules.

El filtre proposat troba cadenes amb una freqüència igual o major a dos que satisfan l'expressió regular següent[34]:

$$((A|N)+ \mid ((A|N)*(N\ P)?)(A|N)*N$$[35]

[34] Aquesta expressió regular cobreix el 97 % del termes presents en diccionaris terminològics (99% si s'accepten les preposicions).

Els candidats a terme de longitud 2 (2 patrons: AN i NA) i longitud 3 (5 patrons: AAN, ANN, NAN, NNN i NPN) són, amb diferència, els més usuals; alguns exemples de candidats a terme amb aquestes diferents estructures:

AN: *linear function, lexical ambiguity*
NN: *regression coeficcients, word sense*
AAN: *Gaussian random variable, lexical conceptual paradigm*
ANN: *cumulative distribution frequency, lexical ambiguity resolution*
NAN: *mean squared error, domain independent set*
NNN: *class probability function, text analysis system*
NPN: *degree of freedom, energy of adsorption*

Aquest algoritme pretén combinar una bona cobertura[36] de la terminologia habitual dels manuals tècnics amb una alta qualitat[37]en l'extracció. En la filosofia mateixa de l'algoritme hi ha una preferència cap a la qualitat en lloc de cap a la cobertura ja que si utilitzés només la restricció gramatical el sistema proposaria molts SN irrellevants. La majoria dels termes rellevants supera la restricció de freqüència.

La selecció dels patrons gramaticals també afecta la qualitat. En particular, si s'admeten les preposicions dintre del patró, s'introdueixen molts candidats, encara que n'hi haurà pocs de vàlids. En conseqüència, la qualitat baixa mentre que la cobertura augmenta, per això l'autor prefereix no incloure-les.

La implementació dels patrons gramaticals també afecta la relació qualitat/cobertura. Hi ha dues aproximacions possibles per atribuir la categoria gramatical a una unitat lingüística: la desambiguació i el filtratge. L'autor descarta la primera opció perquè els desambiguadors no són encara del tot fiables.

El filtratge consisteix a analitzar i lematitzar cada paraula del text i a continuació identificar les seqüències que satisfacin els patrons. Si una paraula no s'identifica com a nom, adjectiu o preposició, es descarta. De cada paraula es conserva només les lectures com a nom, adjectiu o preposició i en aquest ordre de prioritat. La cadena es rebutja si hi ha més d'una paraula que s'identifica com a preposició, o bé no satisfà el patró (per exemple si el patró acaba amb un nom i hi ha més d'una preposició, la paraula que segueix la preposició no és un nom, etc.).

El filtratge té una cobertura com a mínim tan bona com la que s'obtindria amb un analitzador morfològic convencional, malgrat que la qualitat és inferior (per exemple *fixed* s'identifica només con adjectiu *bug fixed,* però també pot ser verb *fixed disk drive*). En contrapartida, la velocitat de processament és notablement superior.

De totes maneres, els autors proposen controlar els patrons, la llista de paraules gramaticals i la freqüència per tal d'ajustar el comportament del sistema a cada tipus de text en particular.

[35] El significat dels símbols utiltzats per aquests autors és el següent: A= adjectiu, N= nom, P= preposició.
[36] Proporció entre els termes vàlids que l'algoritme ha extret i el total de termes del text.
[37] Proporció entre els termes vàlids que l'algoritme ha extret i el total de termes proposats.

El sistema s'ha provat en diferents àrees (metal·lúrgia, enginyeria espacial i energia nuclear) i s'utilitza en centres de traducció d'IBM. En l'article, els autors presenten el resultats de TERMS sobre tres textos tècnics (classificació estadística de patrons, semàntica lèxica i cromatografia). L'avaluació dels resultats l'ha feta el mateix autor de cada article. La cobertura s'ha estimat només per un dels textos i arriba al 71%. La qualitat s'ha avaluat entre el 77% i el 96% de les instàncies.

Valoració

Malgrat que els autors presenten un estudi previ del comportament de les unitats terminològiques, en alguns punts fins i tot amb afirmacions excessivament contundents, el filtre proposat no sembla treure gaire profit d'aquestes anàlisis *a priori* dels termes. A més, s'ha de precisar que aquests tipus de filtres basats en patrons relativament simples no serien tan eficaços si s'apliquessin en altres llengües, com per exemple les llengües romàniques.

2.17 STELLA[38]

Autors:	Jacquin C., Liscouet M.
Publicació:	*Terminology extraction from texts corpora: application to document keeping via Internet*.
Data:	1996
Objectiu:	Eina per facilitar la selecció de documents trobats a Internet
Llengües de treball:	Anglès i alemany

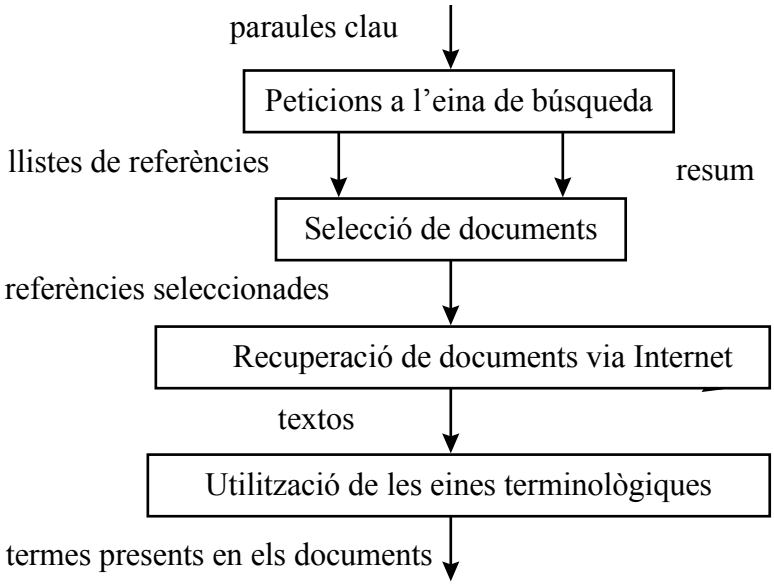
Aquest sistema es presenta fora de l'ordre escollit en aquest treball amb una intencionalitat i raó molt concreta: es tracta d'un sistema cooperatiu. És a dir, dues eines dissenyades i desenvolupades per separat que s'han integrat en un únic sistema. Aquest sistema desenvolupat a l'Institut de Recherche en Informatique de Nantes combina els resultats obtinguts mitjançant el tractament estadístic del sistema ANA (Enguehard/Pantera, 1994) amb l'expansió de termes del sistema FASTR (Jacquemin, 1994).

La idea és crear el nucli inicial de termes no a partir d'un conjunt de termes ja conegut i validat, sinó fent servir una eina de base estadística, com ara ANA. Aquest grup de candidats a terme servirà, després de ser validat, com a nucli inicial o *bootstrap* per FASTR. STELLA està integrat també pel mòdul SQUAL (Morin, 1995) que extreu relacions semàntiques entre les paraules dels candidats a terme. Aquestes relacions s'extreuen amb criteris estadístics i són del tipus *classe de*, *part de* i *causa efecte.*

L'origen d'aquest sistema és el problema que representa buscar informació a Internet ja que les eines estàndard normalment produeixen molt soroll. Per millorar el procés de cerca els autors

[38]*Semantical Terminology: Linguistic and Lexical Adquisition.* Projecte desenvolupat a la Universitat de Nantes. Per a més informació podeu consultar la següent URL:
http://www.sciences.univ-nantes.fr/info/recherche/ln/themeLN.html#Terminologies

proposen eines de detecció de termes com les ja mencionades amb un esquema general de funcionament com el següent :



En aquest sistema, s'accedeix a una eina de cerca estàndard per a la xarxa Internet: LYCOS, on l'usuari escull l'àrea d'interès i el sistema dóna com a resposta una sèrie de referències. A continuació, l'usuari selecciona els documents que *a priori* semblen interessants, STELLA recupera els documents seleccionats i aplica les eines de detecció de termes. L'usuari, a la vista del nombre i tipus de termes detectats, decideix si el document és interessant o no ho és. L'interès radica en el fet de poder seleccionar mots amb una gran càrrega semàntica que reflecteixin per ells sols i de manera molt ràpida el contingut d'un text; així, l'autor no ha de llegir tot el text per decidir si el document és vàlid.

Valoració

Els autors no donen informació sobre els resultats obtinguts amb aquest procediment, factor que dificulta la realització d'una valoració. Malgrat això, la idea de fer treballar conjuntament dos sistemes de naturalesa diversa sembla molt atractiva.

3. Estudi comparatiu

En aquest apartat hem volgut sintetitzar les principals característiques de tots els sistemes analitzats tenint en compte els vuit aspectes següents que creiem que són rellevants a l'hora de dissenyar un nou sistema de detecció d'unitats terminològiques:

- nivells d'informació d'entrada que utilitzen
- estratègies de delimitació de termes
- estratègies de filtratge de termes
- estratègies d'adquisició
- classificació dels termes reconeguts
- interacció amb l'usuari
- resultats obtinguts
- productes comercials

Per a cada un d'aquests criteris hem confeccionat una taula que conté les dades més significatives per tal de facilitar la comparació entre els diferents sistemes.

3.1 Nivells d'informació d'entrada que utilitzen

Una vegada hem analitzat les eines informàtiques que serveixen per extreure de manera més o menys automàtica terminologia a partir de corpus textuals, hem detectat que la majoria d'aquests sistemes utilitzen algun tipus d'informació lingüística, com a mínim una llista de paraules buides per marcar la frontera. La taula següent resumeix de quin nivell d'informació parteixen cada un dels sistemes analitzats:

- no parteixen de cap informació prèvia
- parteixen d'una llista de paraules auxiliars
- usen un analitzador morfològic
- usen un desambiguador
- usen un filtre de categories

Sistema		nivell d'informació d'entrada				
	Nom /autor	cap	llistes de paraules auxiliars	anàlisi morfològica	desambiguador	filtre de categories
1	ANA		X			
2	ATELIER/FX			X	X	
3	AutoLex		X			
4	Blank			X	X	
5	CLARIT			X		
6	Daille			X	X	
7	Drouin			X	X	
8	FASTR			X	X	
9	Heid			X	X	
10	LEXTER			X	X	
11	NEURAL			X	X	
12	NODALIDA-95			X	X	
13	SBIC		X			
14	Termight			X	X	
15	TERMINO			X	X	
16	TERMS			X		X
17	STELLA			X		

Com es pot veure en aquesta taula, gairebé la totalitat dels sistemes utilitzen la combinació d'un analitzador morfològic i d'un desambiguador, tot i que els mateixos autors d'aquests programes coincideixen a considerar el desambiguador com una de les fonts d'error més importants sense donar, però, xifres exactes sobre el grau d'incidència d'aquest fenomen.

3.2 Estratègies de delimitació de termes

Tots els sistemes d'extracció de terminologia en un moment o altre han de decidir l'inici i el final del candidat a terme; és a dir, han de delimitar la possible unitat terminològica. Els programes analitzats es valen d'estratègies diferents per delimitar els termes:

- elements que actuen de frontera de mot
- patrons estructurals
- analitzadors sintàctics
- elements de disposició en el text
- elements tipogràfic
- llista de termes
- desambiguació d'estructures, etc.

El quadre que es presenta a continuació resumeix les diferents opcions adoptades per cada sistema analitzat:

Sistema		delimitació de termes				desambiguació d'estructures	
	Nom /autor	fronteres	patrons	*parser*[39]	altres	aprenentatge	altres
1	ANA				X		-
2	ATELIER/FX				X		-
3	AUTOLEX	X					-
4	Blank	X	X				-
5	CLARIT			X			estadística
6	Daille		X				-
7	Drouin	X					-
8	FASTR		X	X	X		-
9	Heid		X				-
10	LEXTER	X				X	
11	NEURAL		X				-
12	NODALIDA-95				X		-
13	SBIC	X					manual
14	Termight		X				-
15	TERMINO			X			-
16	TERMS		X				-
17	STELLA			X	X		-

[39]En aquest context s'ha d'entendre *parser* com una anàlisi parcial de les frases mai com un intent de donar una anàlisi completa i única de cada frase.

3.3 Estratègies de filtratge de termes

Un punt clau de tot sistema de detecció de termes és el filtratge dels candidats a terme; és a dir, reduir al màxim la llista dels candidats. La taula següent presenta les diferents estratègies usades:

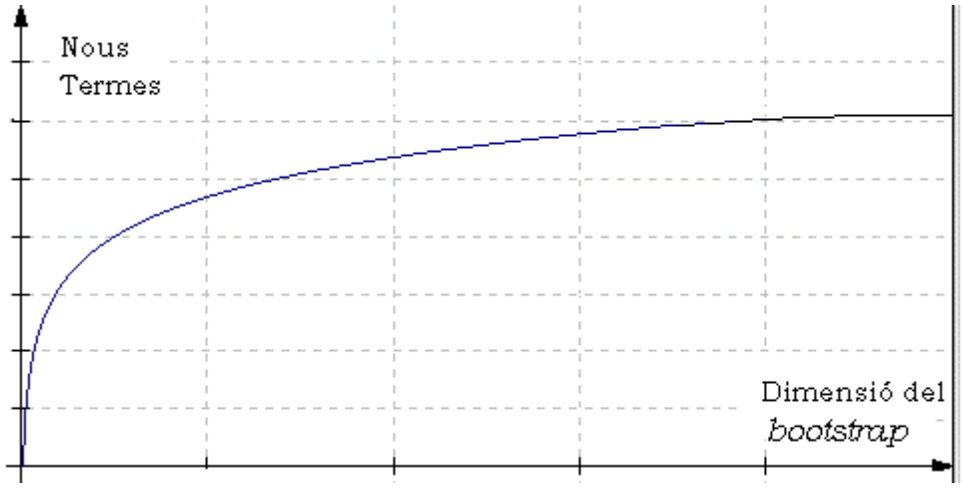
Sistema		filtratge de termes					
	Nom /autor	manual	freqüència [40]	lingüístic	estadístic + lingüístic	lingüístic + estadístic	termes de referència
1	ANA						X
2	ATELIER/FX			?			
3	AUTOLEX	X					
4	Blank		X	X			
5	CLARIT				X	X	
6	Daille					X	
7	Drouin[41]				X		
8	FASTR						X
9	Heid			X			
10	LEXTER			X			
11	NEURAL					X	
12	NODALIDA-95			X			
13	SBIC	X					
14	Termight		X	X			
15	TERMINO			X			
16	TERMS		X	X			
17	STELLA						X

[40] Hem considerat la tècnica de filtratge de termes mitjançant la freqüència com un cas particular, a mig camí entre els mètodes basats en coneixement lingüístics i els mètodes basats en coneixement extralingüístic.
[41] Aquest sistema incorpora també una etapa de postprocessament.

3.4 Estratègies d'adquisició

La gran majoria del sistemes analitzats cada vegada que s'apliquen han de començar el reconeixement des de zero. Només dos sistemes (FASTR, ANA) opten per una estratègia incremental, és a dir, a partir d'un conjunt de termes ja reconeguts el sistema en reconeix de nous. En aquests dos sistemes el mètode de reconeixement és recursiu, encara que a cap dels dos programes es validen els termes detectats abans de fer-los servir en el cicle següent. D'aquesta manera fàcilment un candidat a terme no vàlid pot donar lloc a més termes no vàlids en els cicles posteriors.

El gràfic següent presenta la relació entre el nombre dels termes inicials (*bootstrap*) i la quantitat de termes reconeguts:



Com es pot observar en el gràfic, en una primera fase un nombre petit de termes inicials dóna lloc, després d'algunes iteracions, a la recuperació de molts termes. A mesura que les dimensions del *bootstrap* augmenta el procés tendeix a la "saturació" de tal manera que cada vegada costa més obtenir nous termes. Aquest fenomen es deu bàsicament a les dimensions finites del corpus, la longitud finita dels termes i la definició de candidat a terme adoptada per a aquests sistemes[42].

[42]Un candidat a terme es defineix com una parella (t_i t_{i+1}) amb $0<i<n$ i on t_i és un terme de referència i t_{i+1} s'obté com a una variant de t_i.

3.5 Classificació del termes reconeguts

Alguns dels sistemes estudiats fan una classificació dels termes reconeguts agrupant-los per criteris. D'aquesta manera els termes afins queden més pròxims entre si. Fins i tot hi ha un sistema, FASTR, que intenta deduir una ontologia a partir dels termes reconeguts. Els sistemes que tenen alguna característica en aquest sentit són:

- ANA: crea una xarxa semàntica a partir dels termes detectats
- ATELIER/FX: relaciona totes les paraules d'un text, en principi, només per tal de facilitar-ne la lectura.
- FASTR: crea un graf per relacionar el termes reconeguts. També per alguns termes crea unes ontologies parcials.
- LEXTER: crea una xarxa terminològica mitjançant la descomposició dels termes en nucli i expansió.
- STELLA: extreu relacions semàntiques dels constituents de les unitats terminològiques complexes

3.6 Interacció amb l'usuari

Tots els sistemes que hem descrit (amb l'excepció lògica de CLARIT i STELLA) obtenen un resultat que després ha de ser validat per un terminòleg. En alguns casos el resultat és simplement una llista de candidats a terme mentre que altres sistemes faciliten aquesta tasca de revisió. L'ajuda que ofereixen aquests sistemes sol ser de dos tipus:

- navegació hipertextual: LEXTER i ATELIER/FX
- finestres amb termes i múltiples contextos agrupats: NODALIDA-95, Heid, Termight i TERMINO.

3.7 Resultats obtinguts

El quadre que es presenta a continuació resumeix per a cada sistema el tipus de corpus utilitzat per a les proves i els resultats obtinguts.

Sistema		Corpus de prova			Termes [%]		
	Nom /autor	Àrea	Llengua	Dimen. [K par.]	silenci	precisió	recall
1	ANA	Bricolatge	francès anglès	120 25	?	 75	? ?
2	ATELIER/FX	Medicina	francès	?	?	?	?
3	AUTOLEX	?	?	?	?	?	?
4	Blank	Jurídic (Patents)	alemany	12.000	?	80[43]	?
5	CLARIT[44]	Notícies	anglès	240 Mbytes	-	-	81,6
6	Daille[45]	Telecomunicacions	francès	200 800	?	?	?
7	Drouin	Geomàtica	francès	?	?	?	?
8	FASTR	Medicina	francès	1.560	?	86,7	74,9
9	Heid	Enginyeria	alemany	35	?	?	?
10	LEXTER	Enginyeria	francès	3.250	5	95[46]	?
11	NEURAL	Medicina (oftalmologia)	anglès	55	?	?	70
12	NODALIDA-95	Cosmologia Enginyeria Ind. del Automòbil Premsa	anglès	20		95-98	98,5-100
13	SBIC	Medi ambient	italià	?	?	?	?
14	Termight	Informàtica	anglès	?	?	?	?[47]
15	TERMINO	Medicina	francès	?		72	70-74
16	TERMS	Metal·lúrgia Eng. Espacial Eng. Nuclear Estadística Semàntica Cromatografia	anglès	? ? ? 2,3 6,3 14,9		 71 77 86 96	
17	STELLA	Documents d'Internet	anglès francès	?	?	?	?

En primer lloc, hem de tenir present que la informació del quadre es limita al component monolingüe d'aquest tipus de sistemes. Ara bé, de tots els sistemes indicats en el quadre precedent,

[43] Per a aquells termes de freqüència superior a tres.

[44] El sistema s'ha provat molt intensivament pel que fa a l'eficiència de la indexació, però no en relació a la qualitat del termes extrets.

[45] A Daille i al., 1996 s'indica que una aplicació d'aquest treball sobre un manual de telecomunicacions de 200.000 mots (*The satellite communication handbook*) ha donat una taxa d'encert al voltant de 85%.

[46] Comentari de l'autor.

[47] L'única dada aportada és que el terminòleg extreu 150-200 termes/hora.

els únics que tenen un component multilingüe són Termight (anglès-alemany) i Blank (anglès-alemany-francès). En ambdós casos, en una primera fase el sistema detecta els termes en una llengua i després s'efectua un procés de paral·lelització. Termight realitza la paral·lelització a nivell de paraula, en canvi, en el sistema dissenyat per Blank la paral·lelització es dóna a nivell de frase.

3.8 Productes comercials

De totes les eines presentades en aquest treball, poques s'han arribat a desenvolupar com a productes comercials ja que la majoria són treballs de recerca elaborats en centres d'investigació o bé aplicacions específiques realitzades per companyies privades. Els sistemes que realment es troben en el mercat són els quatre següents:

- AutoLex (100 U$D)
- ATELIER/FX (500 $ canadencs)
- CLARIT (en realitat no és una eina aïllada, sinó que forma part d'un sistema més complex comercialitzat)
- NPTool (no se'n coneix ni el preu ni la disponibilitat real)

4. Conclusions

4.1 En relació als sistemes analitzats

Després d'haver analitzat i valorat la majoria dels sistemes d'extracció de terminologia que s'han dissenyat en els deu últims anys, i tenint en compte que les dades (extretes principalment d'articles, però també de comunicacions, ponències, papers de treball, webs, etc.), no són tan completes com desitjaríem, hem arribat a les conclusions següents:

a) **L'eficàcia** dels sistemes d'extracció presentats és molt **variable**. En la majoria dels casos l'autor no explicita els resultats finals d'una manera clara i quantificable. A l'hora de valorar els resultats també s'ha de tenir en compte que aquests sistemes es proven amb corpus molt petits i altament especialitzats. Aquesta mancança de dades dificulta moltíssim l'avaluació i comparació d'aquests sistemes, tot i que això no impedeix estimar com a interessants les solucions proposades per alguns aspectes determinats.

b) Cap dels sistemes d'extracció d'unitats terminològiques és totalment **satisfactori.** Aquesta afirmació se sostenta bàsicament en dos fets: d'una banda, tots els sistemes produeixen una quantitat massa gran de **silenci** (termes no inclosos entre els candidats a terme) sobretot els de base estadística, i de l'altra, tots generen una quantitat molt elevada de **soroll** (segments discursius o unitats lèxiques inclosos entre els candidats a terme) sobretot els de base lingüística.

c) Atès el soroll que es genera, tots els sistemes d'extracció proposen **llistes de candidats a terme** que, al final del procés, s'han d'acceptar o rebutjar manualment. Consegüentment, podem afirmar que tots els sistemes informàtics d'extracció de terminologia actuals són **semiautomàtics** .

d) La majoria de sistemes s'apliquen a **una sola llengua**, que sol ser el francès o l'anglès. No hi ha cap sistema dissenyat per al català o per al castellà.

e) Com ja hem anticipat, els **corpus** sobre els quals es fan les proves solen ser **petits** (de 2.3 a 12 quiloparaules) i **altament especialitzats** tant pel que fa al tema com al nivell d'especialització. Aquesta característica permet que els patrons i les heurístiques lexicosemàntiques, formals, i morfosintàctiques siguin força precises, però només aplicables a un corpus molt especialitzat temàticament.

f) Tots aquests sistemes avaluats se centren exclusivament en el **sintagma nominal**, cap sistema d'extracció fa referència als **sintagmes verbals.** Aquesta dada respon al fet que el percentatge de sintagmes nominals terminològics en els textos especialitzats és altíssim, si bé aquest percentatge pot variar en funció de la temàtica i del grau d'especialització. El que també és un fet, però, és que en tots els llenguatges d'especialitat hi ha verbs propis, encara que el percentatge sigui molt inferior al dels noms, o combinacions específiques de base verbal. Depenent de l'aplicació de l'extractor de terminologia, per exemple com a suport a la traducció automàtica, s'haurien de tenir en compte els verbs terminològics i la fraseologia.

g) Conseqüentment, cap sistema d'extracció de terminologia fa referència a la delimitació entre col·locacions nominals i unitats terminològiques nominals sintàctiques; ni tampoc a l'extracció de fraseologia verbal.

h) Molts d'aquests sistemes estudiats utilitzen una sèrie de **patrons** morfosintàctics per identificar els termes complexos i, encara que responguin a la majoria d'unitats terminològiques, solen ser molt **reduïts** i, alhora massa poc restrictius: per a l'anglès AN i NN per al francès NA i N prep N, de tal manera que hi ha alguns termes amb altres estructures que no es detecten mai. Els sistemes d'extracció basats només en aquest tipus de tècniques lingüístiques genera massa soroll.

i) Hi ha un consens bastant generalitzat entre els diferents autors a considerar la **freqüència** com un bon indicador que un candidat a terme sigui finalment una unitat terminològica, encara que per si sola **no** sigui del tot **suficient**. Si s'usa només la freqüència, es genera força silenci.

j) Cap dels sistemes analitzats utilitza informació semàntica per reconèixer i delimitar les unitats terminològiques. Cal dir que els sistemes de representació de la semàntica lèxica són actualment pocs, i clarament insuficients. Encara que hi ha alguns sistemes que usen heurístiques amb filtres lexicosemàntics.

k) Cap sistema fa servir a fons les característiques combinatòries pròpies dels termes dels llenguatges d'especialitat lligats a una temàtica. Seria important disposar de més estudis sobre el tipus de restriccions que presenten les unitats terminològiques en relació a:

- camp conceptual
- tipus de text

a) Algunes de les estratègies usades per diferents sistemes que ens semblen particularment interessants són:

- l'ús de regles heurístiques, en relació tant a la unitat terminològica com a allò que no pot ser mai una unitat terminològica
- la construcció de xarxes d'extensions i de nuclis dels termes complexos
- la reutilització de termes ja reconeguts
- l'anàlisi parcial de les frases per obtenir SN potencialment terminològics
- l'extracció de relacions semàntiques entre els termes -o els seus components-
- la importància de les característiques de disposició de les unitats terminològiques en els textos, etc.
- la combinació de més d'una estratègia

Per millorar aquests sistemes d'extracció de terminologia i aconseguir que es redueixi tant el silenci com el soroll que generen, caldria aprofundir principalment en dos tipus d'estudis. D'una banda, caldrien més estudis lingüístics sobre:

- les relacions semàntiques dels termes
- les relacions semàntiques entre els constituents d'una unitat terminològica
- la representació semanticolèxica
- les restriccions de les unitats terminològiques dintre d'un camp especialitzat concret i en un tipus de text concret
- l'estudi de totes les categories gramaticals susceptibles d'ésser termes en els diferents àmbits d'especialitat
- la influència de la funció sintàctica dels sintagmes terminològics en els textos
- les relacions entre els termes i la seva disposició en els textos
- les relacions en llengües diferents dels termes d'una mateixa xarxa conceptual

I d'altra banda, caldria treballar en la idea de sistemes informàtics que:

- alternin de manera més activa els mètodes estadístics amb els lingüístics
- millorin les mesures estadístiques
- combinin més d'una estratègia
- siguin aplicables a més d'una llengua
- millorin les interfícies per afavorir la interacció màquina/usuari

En definitiva, si es vol avançar en el camp de l'extracció automàtica de terminologia, els mètodes estadístics i els mètodes lingüístics s'han d'interaccionar activament. No són, per tant, aproximacions excloents, sinó complementàries. L'objectiu final d'aquestes millores seria reduir al màxim el silenci i el soroll, de tal manera que el procés de buidatge terminològic a partir de corpus textuals especialitzats arribés a ser al més automàtic i precís possible.

4.2 En relació a una proposta d'un sistema integrador de detecció automàtica de terminologia

Una vegada valorats els diferents fonaments de base dels sistemes de detecció semiautomàtica de terminologia i les diverses estratègies que aquests utilitzen, ens decantem per una eina de naturalesa mixta, és a dir, per un sistema que combini els mètodes lingüístics amb els estadístics, o millor dit per un sistema en què les estratègies estadístiques siguin només complementàries de les lingüístiques.

Des d'una posició teòrica, seria desitjable comptar amb un sistema que actués només amb coneixement lingüístic, reproduint la manera com actua teòricament el cervell humà. Si tenim present, però, que en els textos d'especialitat apareixen determinades unitats lèxiques, categories gramaticals, esquemes morfosintàctics i esquemes lexicosemàntics amb molta més freqüència que d'altres, podem deduir que la freqüència d'ús també pot jugar un rol important[48]. És en aquest sentit

[48]Els termes tenen un vessant gramatical i un vessant pragmàtic, i és en els aspectes pragmàtics on trobem la major part de les peculiaritats del lèxic especialitzat: "Els termes, a més de poder ser definits com a unitats gramaticals de triple vessant (formal, conceptual i funcional) que formen part d'un sistema gramatical, són també unitats pragmàtiques de comunicació i de referència i, per tant, apareixen en discursos específics de comunicació" (Cabré, 1992). Des del punt

que considerem l'estadística com un complement del coneixement lingüístic que, en un moment del procés de detecció, pot ajudar, per exemple, a:

- reafirmar la condició de terme d'una unitat
- rebutjar la condició de terme d'una unitat
- detectar fraseologia especialitzada

Els arguments que ens porten a defensar un sistema basat en coneixements lingüístics que es complementi amb tècniques estadístiques són principalment dos; el primer argument, de naturalesa teòrica, fa referència al fet que l'activitat lingüística compta amb coneixements de tipus lèxic, morfològic, sintàctic, semàntic i pragmàtic; en aquest sentit els detectors de terminologia de base lingüística retroalimenten la reflexió sobre el funcionament del llenguatge.

El segon argument, de naturalesa aplicada, es refereix al fet que els mètodes estadístics provoquen una quantitat important de silenci; és a dir, que si s'aplica un procediment estadístic des de l'inici, hi haurà unes paraules que, per la seva baixa freqüència d'ús en els textos seleccionats, no es podran detectar mai. Per això, els mètodes estadístics més eficients són aquells que treballen en corpus textuals molt grans. Els mètodes basats en coneixements lingüístics, en canvi, no generen gairebé silenci, encara que generen una quantitat considerable de soroll que varia en funció del tipus de coneixement lingüístic que utilitzin. Els mètodes lingüístics solen treballar amb corpus reduïts de comprovació.

Encara que totes les aplicacions informàtiques de reconeixement de termes de base lingüística parteixen de corpus textuals especialitzats etiquetats i desambiguats morfològicament[49], no totes utilitzen els mateixos nivells de coneixement lingüístic. Les possibilitats són diverses:

- diccionaris de termes
- diccionaris de paraules auxiliars
- patrons determinats
- elements que marquen fronteres exteriors

Situats en un pla ideal, seria desitjable comptar amb dues eines que actualment no utilitza cap detector de terminologia:

- un analitzador sintàctic
- una base de coneixements terminològics

Un analitzador sintàctic *ideal* permetria:

- treballar amb les assignacions de dependències ja resoltes. D'aquesta manera, s'aconseguiria una millora considerable en la determinació de les unitats terminològiques (UT)
- explorar la relació entre les UTs i el rol sintàctic d'aquestes unitats

de vista de l'ús, doncs, els termes prioritzen unes determinades estructures formals, perseveren en unes determinades categories gramaticals, usen amb més assiduïtat unes combinacions sintàctiques concretes, uns determinats contextos comunicatius, vehiculen una temàtica específica, tenen una funció concreta i s'utilitzen per uns determinats usuaris.

[49]Com ja hem dit en la majoria de sistemes, una part important del soroll que es genera està causat per errors en l'etiquetatge morfològic.

- disminuir la sensibilitat del sistema als errors en la desambiguació de categoria morfològica, ja que l'analitzador podria treballar amb les assignacions de categoria menys segures per resoldre.

S'ha de tenir present que en els textos d'especialitat trobem unitats lèxiques de la llengua general i unitats lèxiques que s'usen en una o més d'una temàtica determinada. Una BCT, entesa com un conjunt de conceptes interrelacionats amb informació lèxica per a cada concepte, permetria activar una xarxa de relacions conceptuals concreta en funció del tipus de text que s'està analitzant. Així, aquesta eina jugaria un rol important en la determinació del candidat a terme i en l'assignació d'una unitat a una àrea determinada.

Ara bé, la realitat és que ni comptem amb una analitzador sintàctic per al català ni per al castellà, ni comptem encara menys amb BCT[50], i les que s'estan dissenyant actualment parteixen d'un detector semiautomàtic de terminologia per començar a alimentar-se (Condamines/Bourigault, 1995). En el fons es tracta d'un circuit tancat: per constituir un BCT necessitem un detector de termes, per dissenyar un detector de termes aniria molt bé poder tenir una base de coneixements terminològics.

No podem perdre de vista, però, que l'objectiu final del treball és el disseny d'un sistema de detecció de candidats a terme que sigui al més fiable possible amb els recursos que són disponibles

Si partim exclusivament de diccionaris especialitzats (seguint la idea de base del FASTR), ens trobarem que els únics termes que podrem detectar són aquells que coincideixen amb el lexicó de referència, totes les variants i termes nous que puguin aparèixer *coordinats* amb els de referència, és a dir, qualsevol terme que tingui un nucli i/o una expansió que coincideixi amb un terme del lexicó. Contràriament, no podrem detectar els termes que no apareixen relacionats d'alguna manera amb els de referència. La utilització d'aquest sistema sembla òptima una vegada s'han reconegut un nombre important de termes, per la qual cosa creiem que és necessari complementar-lo amb un sistema que detecti termes nous deslligats dels termes de referència. La idea de treballar a partir de termes ja reconeguts i acceptats, recollits en un lexicó específic, és no repetir la tasca de detecció cada vegada que s'analitza un text. La creació de nous termes és contínua, especialment en les àrees d'especialitat capdavanteres. Això fa imprescindible una tasca de retroalimentació del sistema, que tindria com a conseqüències immediates una major rapidesa i fiabilitat i, en definitiva, l'eina esdevindria més robusta.

Si partim de patrons preestablerts, només podrem detectar els mots que segueixin exactament els patrons proposats. A més, els sistemes que es basen en estructures preestablertes utilitzen un nombre molt reduït de patrons, amb la qual cosa sempre hi ha termes --encara que el percentatge sigui petit[51]-- que queden exclosos, però si considerem tots els patrons possibles, sense tenir en compte la seva productivitat en un àmbit temàtic, el soroll augmentaria moltíssim.

[50]Fa només uns anys que les BCT es van concebre i encara no n'hi ha cap que funcioni de manera satisfactòria per a cap llengua. L'única base de coneixement lèxic disponible, encara que pel llenguatge general, és WordNet, que els autors defineixen així: "WordNet is an on-line lexical reference system whose design is inspired by current psycolinguistic theories of human lexical memory. English nouns, verbs and adjectives are organised into synonym sets each representing one underlying lexical concept. Different relations link the synonym sets" (Miller i al., 1993).

[51]En un estudi sobre les estructures de les unitats terminològiques polilexemàtiques (UTP) a partir d'un corpus terminogràfic de dret i de medicina (Estopà, 1996) es va comprovar que: "la gran majoria d'UTP es concentren en molt poques estructures (entre el 75% i el 90% de les UTP corresponen entre 3 i 7 estructures diferents), i que,

Si partim només d'elements que marquen frontera de terme, el soroll que genera el sistema és força alt. LEXTER és un dels mètodes que utilitza aquesta estratègia per detectar termes i, encara que té una cobertura del 95% (per tant genera molt poc silenci: 5%), el soroll es situa entre el 40% i el 70%.

El problema de base d'aquests dos sistemes a l'hora de delimitar un terme és que no limiten prou bé el context del candidat a terme i extreuen seqüències sense controlar si es tracta o no d'un sintagma complet. Un analitzador sintàctic podria resoldre aquesta qüestió.

Finalment, tenim l'opció de treballar amb patrons sintàctics i seleccionar les dependències estructurals. Aquesta és l'aproximació utilitzada pel sistema NODALIDA-95 que ha donat resultats molt bons a tenor de l'informe dels autors.

La llista de candidats a terme es pot depurar més si es tenen en compte heurístiques de diferent naturalesa:

- gramaticals (lexicosemàntiques (TERMINO, per posar un exemple), morfològiques (TERMINO, Drouin), sintàctiques (Drouin, CLARIT))

- pragmàtiques (freqüència d'ús (Drouin), disposició del terme dins del text (NODALIDA), tipogràfiques (Drouin, NODALIDA), autonomia discursiva (LEXTER), relació amb els altres termes del mateix paradigma (NOMINO)). És en aquest moment de la depuració dels termes on les tècniques estadístiques poden ajudar a decidir si un terme ho és o no.

Des d'un punt de vista aplicat, la finalitat última és trobar mecanismes de tot tipus que permetin a l'ordinador deduir que una unitat lèxica és una unitat terminològica. Consegüentment, manquen més estudis sobre el funcionament i les característiques dels termes, simples i complexos, en els textos. Treballs basats en la **lingüística de corpus** que tinguin en compte tant el vessant més gramatical de les unitats terminològiques com la vessant pragmàtica. A més a més, també s'hauria d'estudiar l'ordre d'aplicació de les diferents estratègies i el valor de cada una perquè la detecció fos com més fiable millor.

Pel que fa a la interfície i tenint present que l'etapa de validació és ara mateix ineludible, s'ha de proporcionar al terminòleg la major quantitat d'informació en un entorn al més funcional possible que doni un accés fàcil a tota la informació que pugui ajudar a la tasca de validació:

- informació de context amb totes les marques textuals
- relació entre els candidats a UT proposades pel sistema
- UT de referència o ja validades
- informació morfosintàctica dels candidats

contràriament, hi ha una dispersió d'estructures que equivalen a molt poques UTP, pel fet que només les trobem una o dues vegades en tot el corpus (entre el 10% i el 15% de les UTP corresponen a 70/50 estructures diferents)".

Ara per ara, els sistemes encara han de ser aplicables a una llengua concreta i a uns textos d'especialitat molt ben determinats, perquè no comptem amb cap base de coneixements terminològics ni amb cap eina de detecció de xarxes conceptuals independents a les llengües que ens ajudarien a detectar els termes *de* cada una de les especialitat. En definitiva, la concepció i creació d'un sistema d'extracció automàtica de terminologia de base lingüística ens condueix a replantejar-nos teòricament i metodològicament qüestions tan difícils de respondre com:

- què entenem exactament per unitat terminològica?
- què entenem per àrea d'especialitat?
- què entenem per usos especialitzats?
- com podem saber que un terme pertany o s'usa en una àrea d'especialitat determinada?
- com l'ordinador pot detectar que una unitat lèxica pertany o s'usa en una àrea d'especialitat determinada?

La nostra proposta s'inclinaria per adoptar un sistema integrador de diferents estratègies concebudes com a complementàries i que interectuarien entre si.

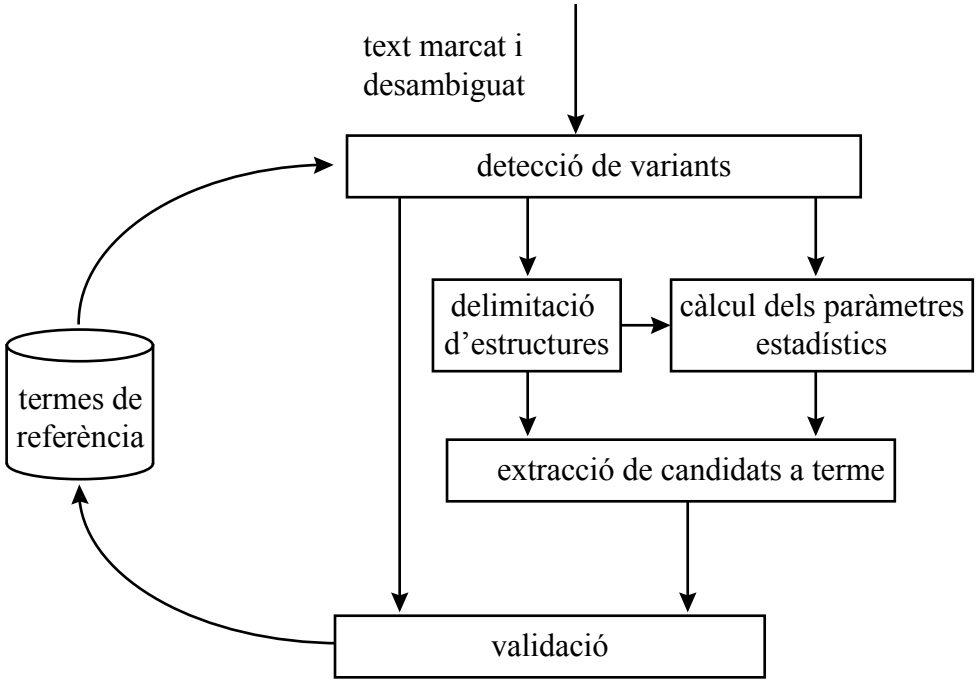
Una seqüència d'accions possible, que, òbviament, s'hauria de comprovar i millorar, podria ser la següent:

- primera fase: detectar els termes (o variants) a partir d'un lexicó especialitzat tenint en compte l'autonomia dels constituents i les relacions paradigmàtiques
- segona fase: aplicar un sistema d'extracció de base lingüística als segments no relacionats amb els de referència
- tercera fase: calcular paràmetres estadístics rellevants: freqüències, grau d'associació,...
- quarta fase: filtrar aquesta llista amb un mòdul de diverses variables complementàries relatives a:
  - paràmetres estadístics
  - la distribució del text sobre el paper[52]: negretes, cursives, subratllats, segments entre parèntesis o guionets, títols, ítems de llistes, cel·les de taules, etc.
  - la funció sintàctica del candidat a terme
  - el valor semàntic de les unitats que formen part del candidat a terme
- cinquena fase: validar els candidats a terme proposats pel sistema amb una interfície al més amigable i eficient possible.
- sisena fase: realimentar el lexicó amb els nous termes validats

La informació d'entrada és text especialitzat en el format en què es troba el corpus de l'IULA, és a dir amb marques estructurals i morfològiques (totalment desambiguat).

En l'esquema que es presenta a continuació es representa gràficament aquesta seqüència d'operacions:

[52] Tota aquesta informació es troba ja codificada en el Corpus Tècnic de l'IULA.


text marcat i
desambiguat
detecció de variants
delimitació
d'estructures
càlcul dels paràmetres
estadístics
termes de
referència
extracció de candidats a terme
validació

5. Bibliografia

1. AHMAD K. i al. (1996). “Engineering Terminology - A case for a linguistically-informed terminology database”. *TKE ‘96: Terminology and Knowledge Engineering*. Berlín: Indeks Verlag. Pàg. 166-178.

2. 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)

3. BENVENISTE, É. (1974). *Problèmes de linguistique générale II.* París : Gallimard.

4. BLANK, I. (1995) "Méthodes pour l'extraction de terminologie bilingue". *Actes de les IVèmes Journées scientifiques du Réseau Lexicologie, Terminologie, traduction.* Lió. [En premsa]

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 reperage de termes complexes dans un texte". *TAL,* 2. Pàg. 105-117.

7. ---. (1995). "Conception et exploitation d'un logiciel d'extraction de termes: problèmes théoriques et méthodologiques". Actes de les *IVèmes Journées scientifiques du Réseau Lexicologie, Terminologie, Traduction.* Lió, [En premsa].

8. ---. (1994). *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes.* París, École des Hautes Études en Sciences Sociales. [Tesi doctoral].

9. ---. (1996). "LEXTER, a Nature Language Processing Tool for Terminology Extraction". Actes del *7th EURALEX International Congress.* Göteborg, [En premsa].

10. BOURIGAULT, D.; CONDAMINES, A. (1995). "Réflexion sur le concept de Base de Connaissances Terminologiques". Actes de les *5èmes Journées Nationales du PRC GDR Intelligence Artificielle* [En premsa].

11. CHURCH, K. W. (1989). “Word association norms, mutual information and lexicography”. Actes del *27th annual meeting of the ACL.* Vancouver. Pàg. 76-83.

12. CONDAMINES, A. (1995). “Terminology: new needs, new perspectives”. *Terminology,* 2, 2. Pàg. 219-238.

13. DAGAN I.; CHURCH K. (1994). "Termight: Identifying and translating technical terminology". Actes de la *Fourth Conference on Applied Natural Language Processing*. Stuttgart.

14. DAILLE, B. (1994). *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Université Paris VII. [Tesi doctoral].

15. DAILLE, B. i al. (1996). "Empirical observation of term variations and principles of their description". *Terminology*, 3, 2 p.197-257.

16. ---. (1995). "Repérage et extraction de terminologie par une approche mixte statistique et linguistique". *TAL,* 36,1-2. Pàg. 101-118.

17. DAVID, S. (1995). *Les Unités nominales polylexicales. Éléments de description et reconnaissance automatique.* Université Denis Diderot. Paris VII. [Tesi doctoral].

18. DAVID, S.; PLANTE. Pàg. (1991). "Le progiciel TERMINO: de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes". Actes del Col.loqui de Montreal *Les industries de la langue: perspectives des années 1990,* 1991, 1. Pàg. 71-88.

19. DROUIN, P. (1996). "Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme". [En premsa]

20. ESTOPÀ, R. (1996). *Les unitats terminològiques polilexemàtiques en els lèxics especialitzats (dret i medicina).* Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada. [Treball de doctorat]

21. ENGUEHARD, C.; PANTERA, L. (1994). "Automatic Natural Acquisition of a Terminology". *Journal of Quantitative Linguistics,* 2, 1. Pàg. 27-32.

22. 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. Pàg. 17-24.

23. FRANTZI, K.; ANANIADOU, S. (1995). "Statistical measures for terminological extraction". Paper de treball del Department of Computing of Manchester Metropolitan University.

24. GUILLET, A. (1990). "Reconnaissance des formes verbales avec un dictionnaire minimal". *Langue française,* 87. Pàg. 52-58.

25. HABERT, B.; NAULLEAU, E.; NAZARENKO, A. (1996). "Symbolic word clustering for medium-size corpora". Actes de *Coling'96*. Pàg. 490-495.

26. HEID, U. i al. (1996). "Term extraction with standard tools for corpus exploration. Experience drom German". *TKE '96: Terminology and Knowledge Engineering*. Berlín: Indeks Verlag. Pàg. 139-150.

27. HULL, D. i al. (1996). "Xerox TREC-5 site report: routing, filtering, NLP and Spanish tracks". Actes del *TREC-5*.

28. JACQUEMIN, C. (1994). "Recycling Terms into a Partial Parser". Actes de la *4th Conference on Applied Natural Language (ANLP'94)*. Stuttgart. Pàg. 113-118.

29. 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. Pàg. 445-458

30. 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. Pàg. 74-83.

31. JUSTESON, J.; KATZ, S. (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text". *Natural Language Engineering,* 1, 1. Pàg. 9-27.

32. KAGEURA, K.; UMINO, B. (1996). "Methods of Automatic Term Recognition". Papers del *National Center for Science Information Systems*. Pàg. 1-22.

33. KARLSSON, F.(1990). "Constraint grammar as a framework for parsing running text". Actes de la *13th International conference on computational linguistic*, vol. 3. Pàg. 168-173.

34. LAURISTON, A. (1994). "Automatic recognition of complex terms: Problems and the TERMINO solution". *Terminology*, 1, 1. Pàg. 147-170.

35. 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. Pàg. 25-43.

36. MILLER G. i al. (1993). *Introduction to WordNet: An On-line Lexical Database.* University of Princeton. [Paper de treball]

37. MORIN, E. (1995). *Acquisition automatique de liens sémantique dans les corpus de textes: Application à l'hyponymie.* Université de Nantes. [Memòria de DEA d'informatique]

38. OTMAN, G. (1991). "Des ambitions et des perfofmances d'un système de dépouillement terminologique assisté par ordinateur". *La banque des mots,* 4. Pàg. 59-96.

39. PERRON, J. (1991). "Présentation du progiciel de dépouillement terminologique assisté par ordinateur: Termino". Actes del col.loqui de Montreal *Industries de la langue: perspectives des années, 1990.* 1991, 2. Pàg. 715-755.

40. PINEIRA-TRESMONTANT, C. (1992). "Reconnaissance automatique des unités syntagmatiques". [Paper de treball] Pàg. 1-13.

41. 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.

42. PUSTEJOVSKY, J. i al. (1993). "Lexical semantic techniques for corpus analysis". *Computational Linguistics*, 19, 2. Pàg. 331-358.

43. SHIEBER, S. N. (1986). “An Introduction to Unification-Based Approaches to grammar”. *CSLI Lecture Notes,* vol 4, Chicago University Press.

44. SMADJA, F. (1991). *Extracting collocations from text. An application : language generation*. Columbia University. Department of Computer Science. [Tesi doctoral]

45. VOUTILAINEN, A. (1993). "NPtool, a detector of english noun phrases". Actes del *Workshop on Very Large Corpora.* Columbus: Ohio State University.

46. ZHAI, C. i al. (1996). “Evaluation of sysntactic prase indexing - CLARIT NLP track report”. Actes del *TREC-5*.