Sistemes de recuperació d'informació i processament del llenguatge natural

Eduard Sosa Jurado
Universitat Pompeu Fabra (Barcelona)
sosa_eduard@fcsc.upf.es

RESUM

El propòsit de l'article és exposar els motius pels quals la recuperació d'informació manté una interacció directa amb el llenguatge. Segons aquesta concepció, es pensa que un Sistema de Recuperació d'Informació que consideri el fenomen lingüístic, pot obtenir un increment en l'efectivitat de recuperació.

S'explica quina pot ser la funció del Processament del Llenguatge Natural aplicat als Sistemes de Recuperació; es presenten els mòduls en què es subdivideix l'anàlisi del llenguatge; es detallen alguns aspectes del llenguatge que són rellevants per la Recuperació d'Informació; s' estableix una relació entre els dos camps i s'explica les limitacions de les tecnologies actuals per obtenir aquesta integració. Finalment s'exposen alguns programes de recerca que s'han desenvolupat.

RESUMEN

El propósito del artículo es exponer los motivos por los que la recuperación de información mantiene una interacción directa con el lenguaje. Según esta concepción, se piensa que un Sistema de Recuperación de Información que considere el fenómeno lingüístico, puede obtener un incremento en la efectividad de recuperación.

Se explica cuál puede ser la función del Procesamiento del Lenguaje Natural aplicado a los sistemas de recuperación; se presentan los módulos en los que se subdivide el análisis del lenguaje; se detallan algunos aspectos del lenguaje que son relevantes para la Recuperación de Información; se establece una relación entre los dos campos y se explican las limitaciones de las tecnologías actuales para obtener esta integración. Finalmente se exponen algunos programas de investigación que se han desarrollado.

Introducció

Tradicionalment l'interès d'investigació en recuperació d'informació s'ha centrat principalment en l'ús de tècniques estadístiques i booleanes. En canvi, per diversos factors que es comenten en l'article, recerques de caire lingüístic han estat menys importants, tant per la comunitat científica com pels sectors comercials. D'una banda, els models tradicionals basats en la combinació d'operadors booleans i càlculs estadístics s'han consolidat pels seus resultats acceptables. D'altra banda, les tècniques de processament del llenguatge natural aplicades a la recuperació d'informació no han verificat de manera generalitzada que augmenti la proporció de documents recuperats rellevants ni tampoc que disminueixi el conjunt de documents recuperats que no són rellevants.

En l'actualitat, els grups de recerca busquen formes alternatives d'explotar la informació que contenen els documents: fent ús de tècniques avançades es pot millorar la manera d'extraure informació dels texts i així mateix es poden obtenir nivells de recuperació més satisfactoris. La primera tasca correspon a l'ús del Processament del Llenguatge Natural (en endavant PLN) en tant que la segona és pròpia de les tècniques de Recuperació d'Informació (en endavant RI). La combinació d'aquests dos camps va ser el propòsit del simposi celebrat el 1990 amb el títol *Text-Based Intelligent Systems* el qual va reunir experts de les dues disciplines.

Els experiments realitzats durant les dècades anteriors han constatat la dificultat d'integrar aquests dos camps. En els seus inicis les fites eren massa ambicioses i en canvi, les tècniques del PLN no eren tant avançades com per poder solucionar les qüestions presentades. Els resultats obtinguts van generar un pessimisme que ha portat alguns autors a plantejar-se les possibilitats d'èxit. A fi de superar les limitacions pròpies d'altres models respecte a la relació que s'estableix entre document i ús del llenguatge, l'objectiu principal del models lin-

güístics és obtenir una representació més detallada i proveir un context on es pugui definir estratègies de recuperació més precises. Els models basats en tècniques estadístiques i booleanes, tot i que són efectius comporten una sèrie d'inconvenients degut a una concepció no gaire ortodoxa en el tractament del llenguatge. Part del soroll que es produeix en les cerques i el resultat de recuperar només un subconjunt del documents rellevants pot tenir una relació directa amb la manca de recursos que contemplin al caràcter ambigu del llenguatge. Sembok i Rijsbergen observen les limitacions dels models tradicionals:

> *«En els models convencionals de recuperació d'informació els documents són representats per una col·lecció no estructurada de descriptors, és a dir, paraules claus. Aquesta representació no és ideal com a indicador del contingut dels documents o consultes en sistemes de RI. Algunes paraules són excessivament específiques, altres massa generals. Els descriptors molt genèrics recuperen documents no rellevants; els descriptors massa estrets deixaran sense recuperar informació útil per a l'usuari.» (Sembok i Rijsbergen, 1990)*

Per tant, amb una consideració del fenomen lingüístic, les tècniques de PLN tenen com a finalitat incrementar la *precision* (precisió) i la *recall* (crida). En la resta de l'article es farà servir «precisió» per fer referència al terme anglès *precision* i «crida» per fer referir-nos a *recall*. La precisió indica la proporció de documents recuperats que són rellevants. La crida és un valor indicador de la proporció dels documents rellevants que són recuperats.

Els programes d'investigació actuals, que malauradament no poden obtenir canvis radicals, al mateix temps que es desenvolupen amb propòsit d'aplicació real pretenen obtenir un pont entre els sistemes de gestió documentals i les tècniques de processament del llenguatge de manera que els avanços en processament del llenguatge puguin ser aplicats als sistemes de recuperació d'informació (en endavant SRI).

1. El concepte de processament del llenguatge natural

El PLN té com a finalitat el reconeixement de la informació expressada en llenguatge humà i la seva aplicació a través de l'ús de sistemes informàtics. L'objectiu principal a què respon és conèixer la manera que el llenguatge es pot fer servir per complir diferents tasques, entre altres la traducció automàtica, la creació d'interfícies basades en llenguatge natural o la recuperació d'informació. Actualment, s'han integrat en el PLN tècniques pròpies de la intel·ligència artificial com són models de representació del coneixement i llenguatges de programació declaratius. Pel que fa a l'estructura dels sistemes de processament del llenguatge generalment es diferencien quatre mòduls, especialitzats en la comprensió de diferents nivells del llenguatge: mòdul morfològic, sintàctic, semàntic i pragmàtic. A més d'aquests mòduls es poden incloure altres nivells com són la informació fonològica o l'anàlisi del discurs.

1.1. Anàlisi morfològica

La funció de l'anàlisi morfològica consisteix a detectar la relació establerta entre les unitats mínimes que formen una paraula. Per exemple, l'analitzador s'encarrega de discriminar l'existència de sufixos i prefixos, la determinació d'una forma verbal, el gènere o el nombre d'un paraula . El lèxic, que manté una estreta relació amb l'anàlisi morfològica, és el conjunt d'informació referent a cada paraula que el sistema fa servir per l'anàlisi. Per cadascuna de les accepcions d'una paraula existeix una entrada lèxica, es a dir, una representació dels seus trets diferencials.

El lèxic, utilitzat pels mòduls morfològic, sintàctic i semàntic, comprèn informació morfològica, la categoria gramatical, irregularitats sintàctiques i representació del seu significat. Segons un procediment comú de representació, les entrades lèxiques amb forma regular només són representades amb l'arrel, de manera que l'analitzador morfològic determina la bona formació de la paraula. En recuperació d'informació, les aplicacions més senzilles que s'han fet de PLN han estat a nivell lèxic, indexant les paraules d'entrada en una forma normalitzada o derivada.

1.2. Anàlisi sintàctica

El mòdul d'anàlisi sintàctica té com a propòsit agrupar en estructures més complexes les unitats que apareixen en la frase tal com sintagmes nominals o verbals. El resultat és la generació d'una estructura formada per les categories sintàctiques de cada una de les unitats lèxiques i sintagmàtiques que formen l'oració. Aquest nivell d'anàlisi s'ha fet servir en sistemes de recuperació d'informació per obtenir una indexació del les paraules per sintagmes nominals. Per exemple, els programes IOTA i CLARIT que s'expliquen en l'últim apartat fan ús d'un tractament dels sintagmes nominals.

1.3. Anàlisi semàntica

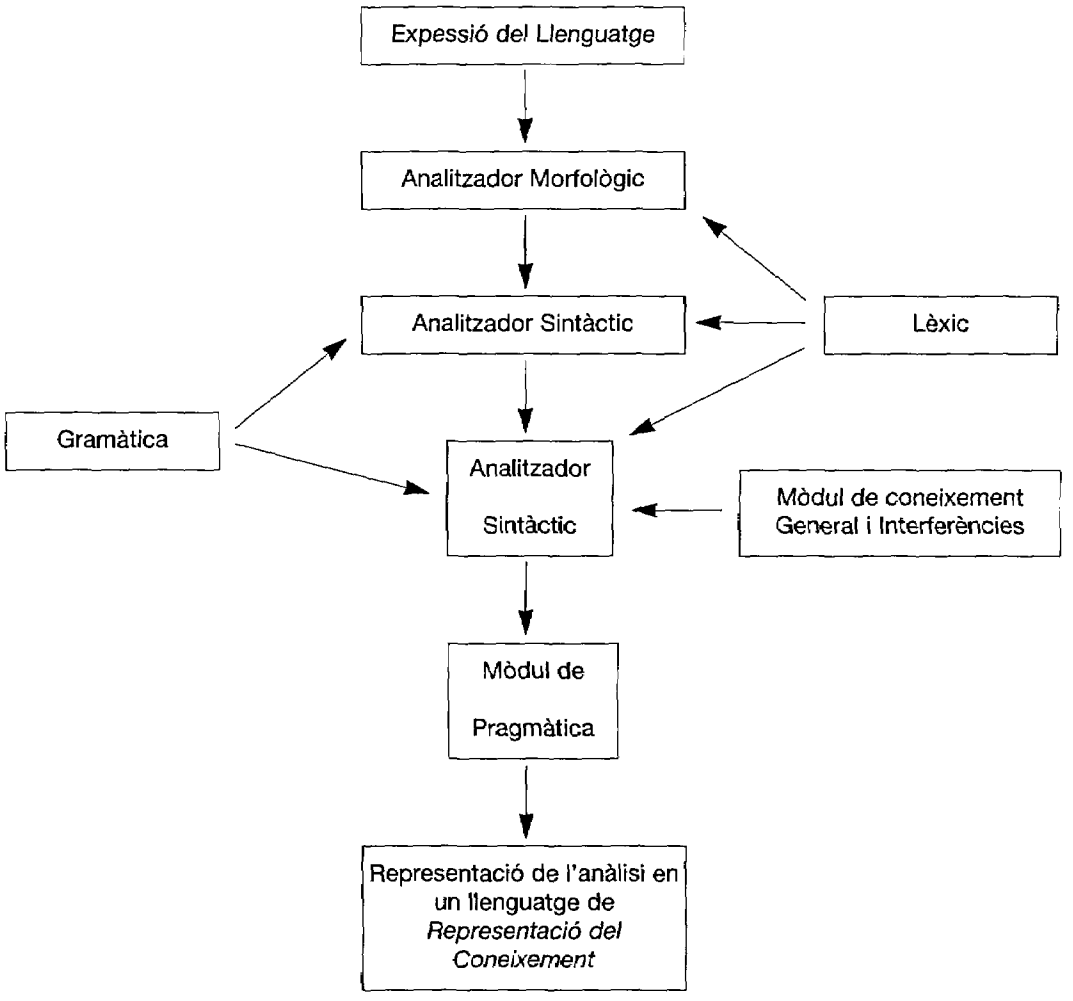
L'anàlisi semàntica té com a finalitat elaborar una representació del significat de les oracions, paràgrafs o documents. A partir de l'anàlisi de les estructures sintàctiques, s'assigna un valor semàntic a cada unitat detectada en el procés anterior. L'anàlisi semàntica s'ha aplicat en SRI per obtenir una representació lògica dels documents i per la creació d'ontologies. Com s'explica més endavant, el programa SILOL proposa un SRI basat en una concepció semàntica.

1.4. Anàlisi pragmàtica

A més dels mòduls anteriors, també pot haver-ne un altre anomenat mòdul d'anàlisi pragmàtica dedicat també a l'assignació del significat. Aquest nivell fa servir informació addicional sobre les relacions que s'estableixen entre el document i el seu entorn d'ús. En l'anàlisi del significat és comú diferenciar el significat que és implícit pròpiament en les paraules del que és afegit pel context. El primer, explicat en el paràgraf anterior, és representat per l'anàlisi semàntica i el segon per l'anàlisi pragmàtica. En SRI aquest nivell ha estat aplicat en dominis de coneixement específics per representar sistemes de recuperació d'informació contextual.

Tot i que aquests nivells d'anàlisi es consideren adequats, no hi ha un consens de com aplicar-los o quin d'ells és el que ha de tenir més importància. Tal com es veurà en l'apartat dedicat a aplicacions desenvolupades, hi ha una diversitat d'aplicacions en les quals els quatre nivells anteriors han estat estudiats.

Representació gràfica dels processos que intervenen en l'anàlisi



2. Aspectes lingüístics rellevants en la representació dels documents.

Les relacions lèxiques i semàntiques juguen un paper important en la comprensió i producció del llenguatge natural i per tant en la representació dels documents. Per aquesta raó, són un element central en les tècniques de representació del coneixement. Les relacions més importants són les de taxonomia, meronímia, polisèmia i sinonímia. A continuació es fa una breu exposició de les seves definicions:

2.1. Taxonomies

Les relacions taxonòmiques associen un conjunt de conceptes específics a un concepte més general amb el qual mantenen una relació conceptual. D'aquesta manera es crea una relació jeràrquica entre diferents nivells de generalitat que poden correspondre a un mateix concepte. Es pot dir que *X és un tipus de Y*, és una relació taxonòmica.

Per exemple :
Cotxe és un tipus de vehicle.
Furgoneta és un tipus de vehicle i
Tot terreny és un tipus de cotxe.

Les taxonomies fan possible la creació d'estructures de representació per inferir informació. Recerques realitzades durant l'última dècada suggereixen que per produir millores en la recuperació cal que d'alguna manera les tècniques aplicades tinguin una certa comprensió del contingut del document i de les consultes.

2.2. Meronímia

Les relacions meronímiques descriuen les relacions part-tot. Aquesta relació considera el grau de diferenciació de les parts respecte a la seva totalitat així com allò que representen les parts respecte al total. La relació meronímica implica que *X té Y* o que *Y és part de X*. En RI es poden fer servir per ampliar o reconduir les cerques.

Per exemple:
Capítol és part de llibre.
Pàgines és part de llibre.
Portada és part de llibre.
Definicions és part de llibre.
i
Pàgines és part de diccionari.
Portada és part de diccionari.
Definicions és part de diccionari.

2.3. Polisèmia

Les relacions polisèmiques descriuen les paraules que poden tenir diversos significats. Aquesta multiplicitat de significats per una mateixa paraula, l'origen dels quals pot ser tant lèxic-semàntic com pragmàtic, implica una reducció en la precisió d'un SRI. Com que les paraules es fan servir per seleccionar documents rellevants, una mateixa paraula pot recuperar entre d'altres documents no rellevants. Per exemple la paraula *organisme* pot fer referència a diferents significats, uns relacionats amb institucions o conjunt de persones i altres relacionats amb els òrgans que formen un ser viu.

2.4. Sinonímia

Dues paraules són sinònimes si tenen un significat semblant. En una llengua no hi ha gaires sinònims absoluts (totalment equivalents), però en canvi si que n'hi ha que tenen una significat pròxim. Pel que fa la RI, el fet de tenir una gamma de paraules per definir el mateix concepte redueix la crida. Per exemple, una cerca sobre organismes podria tenir com a sinònims institució, cos, corporació o entitat.

3. Relació entre Recuperació d'Informació i Processament del Llenguatge Natural

Amb l'aplicació del PLN a la recuperació de documents a text complet, es tracta d'obtenir noves estratègies de cerca no existents en els mètodes tradicionals. La selecció d'informació de forma directa sense les limitacions

que comporta l'ús de descriptors, fa suposar que gràcies a una concepció més lingüística dels documents, es poden produir millors resultats que els obtinguts fins ara. L'increment en la capacitat del processament dels sistemes basats en processament del llenguatge natural i un nou èmfasi sobre la inferència aplicada a la recuperació de texts, suggereix vies de recerca prometedores.

> *«Des d'un punt de vista del PLN, és un repte general el fet d'esbrinar si és aplicable. El fet d'investigar si dades estadístiques i dades no estadístiques es poden combinar apropiadament és un repte específic» (Lewis i Spark Jones, 1996).*

No obstant, l'acceptació d'una nova teoria d'accés a la informació no suposa que es puguin substituir les tècniques aplicades en models tradicionals. Per exemple, si d'aquesta manera és possible expressar una necessitat d'informació amb l'ús de sintagmes nominals (en endavant SN) o frases completes, el resultat de la cerca seria l'obtenció d'una associació conceptual entre el SN de la consulta i els continguts als documents. Com és evident, això no implica necessàriament que aquest document sigui rellevant. El fet que una paraula, sintagma nominal o frase aparegui en una col·lecció de documents no és un indicador suficient per deduir la seva rellevança.

Principalment allò que s'aconsegueix amb la integració de recursos lingüístics és solucionar problemes d'expressivitat del llenguatge: la possibilitat d'expressar un concepte de diferents maneres, comporta una limitació important en la recuperació de documents a text complet. Per tant, la funció pròpia del PLN més aviat està relacionada amb la desambiguació de l'indeterminisme que presenta l'ús del llenguatge. Si amb aquesta concepció dels documents es pot obtenir una representació conceptual del significat, només s'hauria establert un preàmbul desitjable i primordial pel càlcul de la rellevança. Afrontats els fenòmens de caire lingüístic i conceptuals, el càlcul de rellevança depèn d'altres paràmetres que els models actuals tracten d'una manera efectiva.

Un dels avantatges que s'obté amb l'aplicació d'aquestes estratègies té com a finalitat facilitar l'expansió, reorientació o acotació de la cerca. El fet d'expandir allò expressat com a necessitat d'informació, permet associar la consulta amb altres conceptes que si bé no són exactament sinònims mantenen una relació de significat i, d'altra manera formarien part dels documents no recuperats. La capacitat de reorientació o acotació tenen com a finalitat evitar l'indeterminisme del llenguatge i per tant evitar part del soroll que es produeix habitualment.

Fins aquest punt, hem determinat quina pot ser la funció a desenvolupar pel PLN en els sistemes de recuperació d'informació però no s'ha explicat quines són les limitacions i possibilitats reals. Principalment hi ha tres restriccions que posen límits a models basats en el llenguatge: la impossibilitat de construir gramàtiques que puguin tractar tots els fenòmens del llenguatge, els recursos en temps d'ordinador que consumen, i finalment la incapacitat de fer aplicacions que es puguin fer servir de manera universal per treballar amb qualsevol col·lecció de documents.

El primer entrebanc és conseqüència del fet que la investigació en lingüística computacional es troba en estat de desenvolupament: la complexitat implícita en el llenguatge i l'estat emergent de les metodologies en el tractament implica que la seva aplicació encara no sigui prou operativa per un ús comercial. Tal com expliquen Lewis i Spark Jones el fet que les tècniques de PLN funcionin parcialment és un obstacle per la RI:

> *«Estratègies d'anàlisi sintàctica han estat experimentades amb cents de megabytes de text en TR (text retrieval), en canvi l'aplicació d'anàlisi semàntica a gran escala no ha estat demostrat» (Lewis i Spark Jones, 1996).*

En segon lloc, a diferència dels mètodes estadístics, el temps necessari per l'execució d'una consulta i per la representació conceptual del document és elevat, amb la qual cosa es produeix una manca d'operativitat. Com a conseqüència és necessari limitar la grandària dels documents a tractar. El tercer problema té a veure amb la concepció del coneixement humà: el fet de poder obtenir una representació de la informació implícita en un document, no només implica l'ús d'eines lingüístiques sinó també la incorporació d'allò anomenat *coneixement de domini*. En qualsevol domini del coneixement i en la vida quotidiana, per poder entendre els fets i inferir informació es fa servir un conjunt de convencions. Aquest coneixement del món, necessari per obtenir una representació adequada del significat, pot ser representat en dominis de coneixement limitats però no és possible obtenir una representació universal.

4. Aplicacions desenvolupades

En les dues últimes dècades s'han desenvolupat diferents aplicacions basades en models lingüístics. En uns casos s'ha fet servir informació lèxica per millorar els índexs, en altres la informació sintàctica ha permès fer cerques per sintagmes nominals i, projectes més ambiciosos han fet èmfasi en la representació lògica i conceptual dels documents segons l'anàlisi semàntica. A continuació es comenten alguns d'aquests programes.

Des de l'anàlisi sintàctica, la indexació de documents basada en els sintagmes nominals es va fer en el programa IOTA. A efecte d'augmentar l'efectivitat en recuperació, un dels problemes que es va constatar eren la varietats de formes diferents per expressar un concepte en llenguatge natural. Per evitar aquest problema, el programa CLARIT va incorporar un tesaure. El procés de desambiguació en aquest últim consta de dues etapes: primerament tant els documents com la consulta són analitzats per localitzar els sintagmes nominals candidats; en la segona etapa aquests SN són relacionats amb el tesaurus per tal de fer servir sempre la mateixa forma sintàctica en relació a un concepte que s'ha pogut expressar de diferents formes.

El programa Flexible Expert Retrieval of Relevant English Text (FERRET) és un sistema de recuperació d'informació conceptual a text complet. Es fonamenta en una comprensió parcial dels documents amb l'objectiu de millorar la recuperació respecte a les tècniques de cerca per paraules. Fa servir un diccionari electrònic per augmentar el coneixement lèxic i obté una desambiguació tant de les paraules polisèmiques com de les sinònimes.

Els hospitals són un cas paradigmàtic en l'aplicació dels sistemes de recuperació d'informació. El camp mèdic ha mostrat des de fa temps un notable interès per millorar els seus sistemes de recuperació d'informació. La interrelació d'historials és una eina molt valuosa per conèixer com es van resoldre casos semblants en altres pacients. El problema que representa l'aplicació de tècniques del llenguatge natural, és facilitat en aquest camp degut a què es tracta d'un domini restringit (medicina) i l'estil de redacció que es fa servir per elaborar els historials clínics elimina moltes de les ambigüitats del llenguatge. HELIOS, un dels programes de recerca en hospitals, ha estat desenvolupat pel centre d'informàtica hospitalària a la Universitat de Ginebra. L'objectiu principal d'aquest programa és elaborar un entorn en el qual els documents estiguin associats a la seva representació conceptual a fi de poder recuperar els historials mèdics segons el seu contingut. Els mòduls es fan servir són els següents: un analitzador (*Analyser*), un diccionari (*Dictionary Building Tools*) i un sistema de consultes basat en gràfics conceptuals (*Query on Conceptual Graphs*). La finalitat d'aquest tractament és descomposar les frases en fragments significatius per a obtenir una interpretació parcial. A partir d'aquests components es fa una representació del significat amb gràfics conceptuals.

En el programa SILOL, desenvolupat per Sembok i Rijsbergen (Sembok i Rijsbergen, 1989), s'ha aplicat la lògica de primer ordre a la indexació de documents i procés de recuperació. Els resultats constaten una millora en el nivell d'efectivitat en la recuperació, la qual cosa demostra que el tractament amb una teoria semàntica del llenguatge natural i lògica és adequada pels sistemes de recuperació. En la concepció d'aquest model s'entén com a *lògic* les capacitats de fer inferències inductives i deductives, i aplicació de tècniques estadístiques i probabilístiques. Des d'una concepció lingüística s'inclou un tractament sintàctic, semàntic i pragmàtic. Segons els autors, la capacitat lògica és necessària per comparar el document i la necessitat d'informació expressada en llenguatge natural, ja que en altre cas el document per sí mateix no seria prou ric des d'un punt de vista d'implicacions o relacions. Els autors proposen que un model ideal de recuperació d'informació hauria d'estar basat en la unificació de lògica i llenguatge. La fita, a banda d'obtenir una representació lògico-lingüística, és saber com aplicar-ho a la recuperació de documents.

El Centre National de la Recherche Scientifique de Paris ha desenvolupat una estratègia per obtenri una representació formal del contingut semàntic en documentas amb llenguatge natural. Per representar la informació fan servir el *Narrative Knowledge Representation Language (NKRL)*: es tracta d'un llenguatge de representació del coneixement dissenyat per representar el contingut semàntic de documents en llenguatge natural. Un prototip complet del NKRL s'ha implementat en COMMON LISP en el marc del projecte NOMOS.

Conclusions

En aquest article s'han presentat algunes idees sobre la combinació d' estratègies de recuperació amb tècniques de processament del llenguatge natural. La interacció d'aquests dos camps pot proporcionar resultats millors en la representació dels documents i de les consultes. Degut a la dependència que té la recuperació d'informació de l'ús del llenguatge, l'estudi en comú de les dues disciplines proveirà en el futur noves línies d'investigació.

Bibliografia

ALLEN, J. (1995). *Natural language understanding.* Redwood City: Benjamin/Cummings.

BACH, E.W. (1989). *Informal lectures on formal semantics.* Albany: State University of New York Press.

CHIARAMELLA, Y.; *et al.*, (1987). «IOTA: a full text information retrieval system». En: *Proceedings of the ACM Conference on Research and Development in Information Retrieval.* Pisa: F. Rabitti. p. 207-213.

CHIERCHIA, G. (1990). *Meaning and grammar: an introduction to semantics*. Cambridge: MIT Press.

EVANS D.A. (1990). «Concept management in text via natural language processing: the CLARIT approach». En: A *Working Notes for the AAAI Spring Symposium on Text-Based Intelligent Systems.* Stanford.

GAZDAR, G.; MELLISH C. (1989). *Natural language processing in Prolog an introduction to computational linguistics* . Wokingham: Addison-Wesley.

GRISHMAN, R. (1991). *Introducción a la lingüística computacional.* Madrid: Visor.

JACOBS P. (1992). *Text-Based intelligent systems*. New Jersey: Lawrence Erlbaum.

LEWIS D. ; SPARCK JONES, K. (1996). «Natural Language Processing for Information Retrieval» *Comunications of the ACM.* Vol. 39, n°1, January.

MCENERY, T. (1992). *Computational linguistics: a handbook & toolbox for natural language processing.* Wilmslow: Sigma.

REICHGELT, H. (1991). *Knowledge representation: an AI perspective*. Norwood: Ablex.

SEMBOK, T. M. T; VAN RIJSBERGEN. (1990). «SILOL: a simple logical-linguistic document retrieval system». *Information Processing & Management*. Vol. 26, n° 1, p. 111-134.

ZARRI G.P. (1994). «A rule-based approach to the semantic parsing of natural language documents»: *Int.Journal of Applied Expert Systems*. N°1, p. 39-53.