Estratègies estadístiques aplicades a l'extracció automàtica de terminologia

Mercè Vàzquez Garcia

TESI DOCTORAL UPF / ANY 2014

DIRECTOR DE LA TESI

Dr. Antoni Oliver Gonzàlez
Estudis d'Arts i Humanitats
Universitat Oberta de Catalunya

A la meva família

Agraïments

Agraeixo sincerament l'oportunitat que he tingut d'explorar el terreny de la terminologia computacional gràcies al meu director de tesi, el doctor Antoni Oliver Gonzàlez, professor de la Universitat Oberta de Catalunya. Les seves aportacions constants han estat de gran ajuda en la preparació d'aquest treball.

Aixi mateix, agraeixo al doctor Antoni Badia Cardús i a tots els membres del Departament de Traducció i Ciències del Llenguatge de la Universitat Pompeu Fabra la possibilitat que m'han ofert d'explorar les diverses àrees de la ciència cognitiva gràcies al programa de doctorat interuniversitari i interdisciplinari en Ciència Cognitiva i Llenguatge.

Dono les gràcies a l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra i al centre de terminologia Termcat per la bona disposició que han tingut a cedir-me les dades que m'han permès dur a terme aquesta recerca.

Agraeixo molt especialment el suport que he rebut dels meus companys dels Estudis de Ciències de la Informació i la Comunicació de la Universitat Oberta de Catalunya, especialment a Lluís Pastor com a director dels Estudis i als membres del grup de recerca KIMO per les indicacions i orientacions que han aportat al treball. I també als professors Julio Meneses, Julià Minguillón i Angel A. Juan, per l'ajuda que m'han ofert en diferents etapes d'aquest treball.

I al Josep, que sempre hi és.

Abstract

Terminology is found in all areas of knowledge. Due to the use of technology in the different ambits of society, new terms are being created and distributed very quickly and efficiently. Over recent decades, automatic term extraction methods have been developed based on linguistic analysis, statistical strategies and a combination of the two to aid manual extraction. However, these automatic methods tend to produce large numbers of term candidates, which makes manual candidate validation tasks more difficult. This thesis presents an algorithm that uses the terms from a specialist area to detect new terms (token slot recognition method) and lexical association measures to overcome these limitations. It also shows the level of performance offered by the combination of statistical strategies analysed. The token slot recognition method extracts candidates that are more likely to be terms and is able to process corpora in different languages and specialist areas. The research also confirms that lexical association measures place terms in the top positions in lists of candidates and, as a result, aid the final manual candidate validation tasks. In conclusion, the combination of statistical strategies analysed offers flexibility when identifying and validating the terms present in a specialist corpus, which raises the possibility of integrating them into a term extraction tool.

Resum

La terminologia és present en totes les àrees de coneixement. Amb l'ús de la tecnologia en els diferents àmbits de la societat, la creació i difusió de nous termes és molt ràpida i efectiva. En les darreres dècades s'han desenvolupat mètodes d'extracció automàtica de termes basats en anàlisi lingüística, estratègies estadístiques i una combinació de les dues modalitats per a facilitar el buidatge manual d'aquestes unitats, però aquests mètodes tendeixen a extreure un alt nombre de candidats a terme, i aquest fet dificulta la validació manual dels candidats. En aquesta tesi hem dissenyat un algorisme que aprofita els termes presents en un àmbit d'especialitat per a detectar-ne de nous (mètode *token slot recognition*) i fa ús de mesures d'associació lèxica per a poder resoldre aquesta limitació. El treball presenta el nivell de rendibilitat que ofereix la combinació d'estratègies estadístiques analitzades. Hem observat que el mètode *token slot recognition* extreu els candidats que tenen més probabilitat de ser terminològics i té capacitat per a processar corpus en diferents llengües i àmbits d'especialitat. La nostra recerca també confirma que les mesures d'associació lèxica situen els termes en les posicions inicials d'una llista de candidats i, en conseqüència, faciliten la tasca de validació manual final dels candidats. Com a conclusió, la combinació d'estratègies analitzades ofereix flexibilitat a l'hora d'identificar i validar els termes presents en corpus d'especialitat, fet que permet plantejar la seva integració en una eina d'extracció de terminologia.

Sumari

Índex de figures

Índex de taules

Capítol 1

INTRODUCCIÓ

La terminologia, entesa com a conjunt de denominacions que pertanyen a una llengua d'especialitat, és present en totes les àrees de coneixement (ISO, 2000; UNE, 2009). Les unitats terminològiques, d'acord amb el caràcter multidisciplinari de la matèria, són unitats multidimensionals, alhora lingüístiques, cognitives i comunicatives. Els termes, des d'un punt de vista lingüístic, són unitats que comparteixen moltes característiques amb les unitats de la llengua general, tenint en compte que la comunicació especialitzada té un estatus que no és completament aliè al que té la comunicació general (Cabré, 1999a, p. 130).

Així mateix, els termes són unitats lèxiques, activades singularment per les seves condicions pragmàtiques d'adequació a un tipus de comunicació. Aquestes unitats estan formades per forma o denominació i significat o contingut. La forma té les característiques generals de la unitat i el contingut se singularitza fent una selecció de trets adequats a cada tipus de situació, els quals són determinats per l'àmbit, el tema, la perspectiva com s'aborda el tema, el tipus de text, l'emissor, el destinatari i la situació. El valor d'un terme és definit pel lloc que ocupa en l'estructuració conceptual d'una matèria d'acord amb uns determinats criteris. Així, els termes no pertanyen a un àmbit, sinó que són usats en un àmbit amb un valor específic (Cabré, 1999a, p. 132-133).

Amb l'ús creixent de la tecnologia en tots els àmbits de la societat, la creació de noves unitats terminològiques i la corresponent difusió cada vegada és més ràpida i efectiva. Aquest fet ha motivat que en les darreres tres dècades s'incorporin processos automàtics aplicats a la detecció d'unitats terminològiques presents en les diferents àrees de coneixement. L'interès per automatitzar la identificació i extracció posterior d'unitats terminològiques va sorgir paral·lelament a l'aparició dels sistemes operatius emprats per a processar dades. D'aquesta manera, s'aprofita la capacitat de processament que tenen els ordinadors per a dissenyar mètodes d'extracció automàtica de terminologia a fi de superar la limitació d'abast que té el buidatge manual d'unitats terminològiques de corpus especialitzats.

> "El buidatge terminològic és una operació que consisteix a extreure dels corpus de buidatge aquells segments que es consideren termes propis d'un camp d'especialitat del qual s'elabora la terminologia [...]. La primera acció d'aquesta fase de buidatge consisteix a reconèixer en els textos dels corpus de buidatge els segments lingüístics que corresponen a un concepte de l'àrea especialitzada i a delimitarlos. Un especialista en una determinada matèria reconeix amb més facilitat que no pas un llec en la matèria aquests segments terminològics que representen conceptes de la seva disciplina." (Cabré, 1992, p. 273)

La preparació del buidatge manual de termes d'un corpus especialitzat és una tasca llarga, feixuga, repetitiva, que té el risc de ser poc sistemàtica i subjectiva, molt costosa en termes econòmics i limitada per l'abast actual de la informació. Per aquest motiu, l'aparició de sistemes de detecció automàtica de termes representa un avenç significatiu respecte del buidatge manual, ja que permeten el tractament de grans corpus textuals i l'aplicació de criteris de selecció de termes d'una manera sistemàtica.

De manera general, les diferències que s'observen en els resultats obtinguts del buidatge manual i el buidatge automàtic basat en informació lingüística que es duu a terme en corpus amb diferents nivells d'especialització, se centren en el nombre i el tipus de candidats a terme identificats: amb el buidatge manual s'identifica un major nombre d'unitats terminològiques que amb el buidatge automàtic i, alhora, amb el buidatge automàtic s'obtenen candidats a terme que un especialista no reconeix com a especialitzats. En aquest sentit, els sistemes de buidatge automàtic introdueixen silenci pel fet de no reconèixer unitats que són terminològiques i també soroll perquè extreuen candidats a terme que no tenen un caràcter especialitzat. De la mateixa manera, s'ha observat que els sistemes d'extracció de candidats a terme basats en patrons lingüístics generen menys soroll com més alt és el nivell d'especialització d'un text i també que ofereixen una major variació morfològica (Estopà, 1999). La diferència que hi ha entre els resultats obtinguts amb el buidatge manual i el buidatge automàtic incideix directament en la rendibilitat de l'extracció de termes que ofereixen els sistemes automàtics i també en el tipus de treball terminològic que es pugui plantejar en col·laboració amb els especialistes, molt avesats a treballar amb llistes de termes convenientment seleccionades.

En la implementació de sistemes d'extracció automàtica de terminologia s'han de considerar certes limitacions a l'hora de recuperar la totalitat de termes d'un corpus, ja sigui per problemes amb la desambiguació morfosintàctica o la detecció d'unitats complexes i anàfores discursives (Estopà *et al.*, 1998; Estopà, 2007); els resultats obtinguts automàticament mantenen un cert distanciament amb els que s'obtenen amb el buidatge manual, i també s'observa una manca d'adequació dels resultats obtinguts amb les necessitats que tenen els potencials usuaris dels termes (Vivaldi i Rodríguez, 2007; Estopà, 2009).

Així mateix, els sistemes d'extracció automàtica s'han d'adaptar a la tipologia lingüística pròpia de cada llengua per a poder ser eficients. Aquests sistemes han de ser capaços de processar adequadament els trets lingüístics que tenen les *llengües aïllants* com ara el xinès, el tailandès, el vietna-

mita o el hawaià, que compten amb una morfologia nul·la o molt reduïda, amb paraules invariables (sense flexió), amb una relació biunívoca entre paraules i morfemes, i amb unes relacions gramaticals que es manifesten per mitjà de l'ordre de les paraules. De la mateixa manera, les *llengües aglutinants* com ara el basc, l'hongarès, el turc, el finès, el japonès o el suahili compten amb una morfologia molt rica, més que no pas la de les llengües flexives, amb paraules variables (amb flexió), disposen d'una relació biunívoca entre morfemes i morfs i algunes vegades poden presentar al·lomorfisme. Les *llengües flexionals* com l'àrab, l'hebreu i la majoria de llengües indoeuropees expressen les funcions morfosintàctiques dels mots per mitjà de la flexió, motiu pel qual compten amb una morfologia rica, amb paraules variables (amb flexió), amb una relació no biunívoca entre morfemes i morfs, i poden presentar al·lomorfisme. I les *llengües incorporants*, que corresponen a les llengües ameríndies, amb models de mots que són molt complexos, com la inclusió del complement del verb en la mateixa paraula. La varietat tipològica queda reflectida en els sistemes d'extracció automàtica de termes per mitjà de l'adaptació dels mètodes lingüístics i estadístics a les característiques pròpies de cada llengua. En la nostra proposta experimental presentem una combinació d'estratègies estadístiques aplicades al tractament de llengües flexionals (català, espanyol, anglès i francès). Convé assenyalar que aquest plantejament experimental no seria adequat ni satisfactori per al tractament de llengües amb una altra tipologia lingüística, com ara les llengües aglutinants, per al processament de les quals cal fer servir mètodes lingüístics.

La incorporació de sistemes automàtics per a la detecció de termes ha consolidat el paper de la terminologia computacional en la gestió terminològica i ha fet avançar l'estudi i aplicació de diferents estratègies destinades a la detecció d'unitats terminològiques. La publicació de diversos estudis exhaustius que fan referència a la rendibilitat i al nivell d'aprofitament de resultats que actualment ofereixen aquests sistemes, aixi ho constata (Kageura i Umino, 1996; Cabré *et al.*, 2001; Vivaldi, 2001; Pazienza *et al.*, 2005; Vivaldi i Rodríguez, 2007; Korkontzelos *et al.*, 2008; Loginova *et al.*, 2012).

Delimitació del tema i antecedents

En el present treball implementem un procés recursiu d'extracció automàtica de candidats a terme en corpus especialitzats que combina l'ús de *tokens* terminològics, és a dir, seqüències contínues de caràcters presents específicament en els termes, i mesures d'associació lèxica. Aquesta proposta experimental té com a objectiu final avaluar la rendibilitat que ofereix una combinació d'estratègies estadístiques aplicades a l'extracció automàtica de termes. Per a avaluar els resultats emprem els termes seleccionats i validats manualment per terminòlegs dels corpus especialitzats que utilitzem per a implementar el nostre algorisme.

El projecte de recerca *Anàlisi de tècniques estadístiques d'extracció automàtica de termes* (Vàzquez i Oliver, 2007) esdevé l'antecedent immediat d'aquest treball. L'estudi comparatiu de dotze mesures d'associació lèxica[1]aplicades a l'extracció automàtica de termes del corpus anglès de telecomunicacions procedent del projecte Crater (*Corpus Resources and Terminology Extraction*)[2] ens va permetre constatar que el càlcul estadístic de freqüència extreu un major nombre de termes que la resta de mesures avaluades. Els resultats obtinguts mostraren que, de les dotze mesures analitzades, solament cinc (log-likelihood, informació mútua, *t* de Student, Poisson-stirling) van extreure un nombre similar de termes de referència que el càlcul freqüència, i també que *t* de Student fou l'única mesura que aconseguí extreure un nombre de termes una mica superior al de freqüència en les posicions inicials dels resultats.

[1]El càlcul estadístic de freqüència, el coeficient dice, la prova Fishers twotailed, el coeficient jaccard, la ràtio log-likelihood, la mesura informació mútua, la mesura pointwise mutual information, la ràtio odds, la prova khi quadrat de Pearson, la prova *t* de Student, la mesura Poisson-stirling, el coeficient PHI.

[2]El corpus Crater és el resultat de la col·laboració entre la Universitat de Lancaster i la Universitat Autònoma de Madrid, consta de tres llengües (anglès, francès i espanyol) i ha estat elaborat a partir d'una important col·lecció de textos tècnics procedents de la International Telecommunications Union (ITU).

A més, l'avaluació dels resultats obtinguts amb les dotze mesures d'associació lèxica feta per diferents especialistes va evidenciar un nivell d'acord baix en la selecció dels termes. Tal com posen de manifest Vivaldi i Rodríguez (2007) en el seu estudi sobre sistemes d'extracció de termes, aquest baix nivell d'acord es deu principalment a la dificultat que hi ha entre els especialistes sobre què ha de ser considerat un terme. Per a constatar aquesta manca d'acord, mostren els resultats obtinguts d'avaluar els termes presents en un subcorpus de textos universitaris per part de tres especialistes. Així, els termes escollits per tots tres especialistes arriba a un nivell d'acord del 37%, el nombre de termes triat per dos dels especialistes correspon a un 26% d'acord, i un 37% de termes és escollit per un sol especialista. Per contra, algunes unitats lèxiques identificades com a termes no van ser escollides per cap dels especialistes. En el cas del nostre treball vam avaluar manualment els primers dos-cents candidats corresponents a les cinc mesures que van obtenir un major nombre de termes en l'estudi comparatiu de dotze mesures d'associació lèxica (freqüència, log-likelihood, informació mútua, *t* de Student, Poisson-stirling). La selecció manual de termes la van fer cinc especialistes –dos enginyers i tres lingüistes–, els quals van revisar els candidats tenint en compte si aquests corresponien a termes propis de l'àmbit de les telecomunicacions, si eren termes que pertanyien a altres àmbits i que en l'àmbit de les telecomunicacions tenien un caràcter especialitzat o si eren paraules de la llengua general que, pel fet de ser usades en el corpus, van esdevenir termes. La revisió manual feta va constatar un nivell d'acord entre els cinc especialistes d'un 48% en la selecció dels termes, que correspon a un total de 96 termes; un total de quatre especialistes va coincidir en la selecció de 56 termes, amb un nivell d'acord del 28%; tres especialistes van seleccionar 31 termes, amb un 15,5% d'acord; dos especialistes van identificar 22 termes, amb un acord d'1,21%, i un sol especialista va triar 23 termes. En la taula 1.1 recollim la distribució dels resultats fruit de la revisió manual dels candidats a terme.

Taula 1.1: Resultat de l'avaluació feta per especialistes.

Avaluació	Avaluador1	Avaluador2	Avaluador3	Avaluador4	Avaluador5
Terme	176	156	183	173	144
No terme	24	44	17	27	56
Total	200	200	200	200	200

Els termes que van ser seleccionats per consens entre tots els especialistes i també els que van ser triats per quatre especialistes, juntament amb els termes de referència que teníem disponibles en començar a preparar la part experimental del treball, van servir per a identificar quina mesura de les cinc avaluades permetia obtenir un major nombre de termes. Els resultats obtinguts (taula 1.2) van confirmar que el càlcul de freqüència predomina per sobre de la resta de mesures i van consolidar les dades extretes de l'anàlisi comparativa de les dotze mesures d'associació lèxica.

Taula 1.2: Resultats de l'avaluació manual de termes.

Mesures d'associació lèxica	**Nombre de termes**
Càlcul de freqüència	126
t de Student	124
Mesura Poisson-stirling	115
Mesura informació mútua	113
Ràtio log-likelihood	112

Així mateix, amb l'avaluació manual vam constatar la complexitat que representa disposar d'especialistes per a avaluar el nombre de termes que havíem extret amb les mesures d'associació lèxica, ja que en aquests casos cal comptar amb diversos especialistes d'un àmbit per a contrastar els resultats obtinguts, convé especificar ben bé quin tipus d'unitat ha de seleccionar l'especialista, cal disposar de temps per a fer aquesta revisió

manual i és aconsellable de preveure el nombre de candidats que s'ha de revisar.

En els darrers anys les mesures d'associació lèxica han estat aplicades a l'extracció automàtica de terminologia amb l'objectiu de millorar la precisió en la tasca d'extracció de termes d'un corpus. Així, en Daille (1997) s'incorporen les mesures informació mútua, log-likelihood i freqüència aplicades tasca d'extracció de termes combinant estratègies lingüístiques i estadístiques. Com a conclusió de l'estudi s'indica que la mesura informació mútua no ofereix bons resultats, fet que s'atribueix a l'ús de filtres lingüístics; la freqüència obté bons resultats en la identificació de termes, i hi ha preferència per triar la mesura log-likelihood en lloc de freqüència per la millor puntuació obtinguda.

En Pazienza *et al.* (2005) es fa un altre estudi detallat de diverses mesures d'associació lèxica aplicades a l'extracció de terminologia: freqüència, *t* de Student, informació mútua, dice, log-likelihood. Els resultats, per mitjà d'histogrames, no mostren un comportament ideal en cap mesura. Quant a la precisió i cobertura dels resultats, s'observa que hi ha un grup de mesures que inicialment tenen una precisió alta i que a partir d'un determinat moment baixa considerablement (freqüència, *t* de Student, log-likelihood). També s'observa que hi ha un altre grup de mesures que inicialment tenen una precisió baixa i que després va pujant (informació mútua, Dice). Com a conclusió de l'estudi s'indica que les mesures d'importància d'associació (*t* de Student i log-likelihood) funcionen més bé que no pas les de força d'associació (informació mútua, dice).

I en Boulaknadel *et al.* (2008) s'implementen les mesures log-likelihood, *t* de Student i informació mútua en candidats bigrams per tal d'extreure terminologia de textos àrabs. La mesura Log-likelihood assoleix un 85% de precisió; *t* de Student, un 57%, i informació mútua un 26%.

Les mesures d'associació lèxica també són implementades amb l'objectiu d'identificar col·locacions, denominació que fou creada per Firth (1957)

per a referir-se a combinacions de paraules característiques o habituals i que Choueka (1988) va definir en sentit ampli com a "syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components" (*inspirar confiança*, *afirmació rotunda*, *crua realitat*). La tipologia de les col·locacions és força variada, i sovint correspon al tipus de relació que s'estableix entre els elements que hi ocorren, que pot fer referència a combinacions del tipus N Adj (*lluita acarnissada*, *malaltia contagiosa*); expressions idiomàtiques, en les quals cap de les paraules de la col·locació contribueix al sentit general de la mateixa col·locació (*de pa sucat amb oli* <mediocre, de poc valor>); col·locacions estretes, en les quals com a mínim una de les paraules contribueix al sentit de la col·locació (*eixugar la butxaca* (a algú) <deixar-lo sense diners>), o frases fixades, en les quals totes les paraules contribueixen al sentit de la col·locació (*pa amb tomàquet*). Les col·locacions es caracteritzen per tenir tres propietats: *composicionalitat limitada*, pel fet que és un conjunt de paraules que tenen un determinat sentit per anar juntes i que representen un concepte únic, i si aquestes mateixes paraules es troben per separat, llavors representen un concepte diferent; *substitució limitada*, ja que les paraules d'una col·locació no poden ser substituïdes per unes altres de semànticament semblants i mantenir el mateix sentit; *modificació limitada*, referit al fet que una col·locació tampoc no pot ser modificada amb recursos lèxics addicionals (McInnes, 2004; Wermter i Hahn, 2004; Evert, 2005).

Tenint en compte que les paraules que formen part d'una col·locació tendeixen a aparèixer juntes amb més freqüència que altres combinacions de paraules, les mesures d'associació lèxica permeten identificar-les d'una manera automàtica (McInnes, 2004), i així ser utilitzades per a la construcció de lexicons, recuperació d'informació i traducció automàtica. En aquest sentit, en el treball de Krenn i Evert (2001) s'analitza la rendibilitat de les mesures d'associació lèxica informació mútua, dice, khi quadrat de Pearson, *t* de Student i log-likelihood en la detecció de col·locacions formades per verbs. Els resultats s'analitzen segons el percentatge de precisió que obtenen i són contrastats amb els de freqüència. En aquesta proposta experimental es constata que no hi ha cap mesura que obtingui

millors resultats que la freqüència. En la mateixa línia d'estudi, disposem de l'avaluació empírica de les mesures d'associació lèxica corresponents a freqüència, *t* de Student, log-likelihood i khi quadrat de Pearson feta per Evert i Krenn (2005) i destinada a l'extracció de col·locacions lèxiques d'un corpus. Concretament l'estudi es basa en l'anàlisi d'un parell d'estructures sintàctiques en alemany i indica que les mesures *t* de Student i log-likelihood obtenen millors resultats que no pas la freqüència i khi quadrat de Pearson.

En definitiva, tant els resultats obtinguts en el treball de recerca previ a aquesta tesi com els que mostren les publicacions de referència en aquest àmbit constaten la rellevància del càlcul estadístic de freqüència i de les mesures *t* de Student i log-likelihood aplicades a la tasca d'extracció automàtica de terminologia i a la detecció de col·locacions. En el treball que presentem, fem una proposta experimental d'extracció automàtica de termes a fi de millorar la rendibilitat que ofereix el càlcul estadístic de freqüència i analitzem la capacitat que tenen les mesures d'associació lèxica de millorar la tasca de validació manual de candidats a terme.

1.1 Objectius

Els diversos mètodes d'extracció automàtica de terminologia que han estat desenvolupats en els darrers trenta anys mostren la necessitat i l'interès que hi ha d'automatitzar la tasca d'identificació dels termes en un corpus especialitzat. A continuació presentem les principals premisses que tenim en compte en aquest treball basant-nos en els mètodes d'extracció automàtica de termes publicats en treballs previs (capítol 3) i plantegem els objectius de la present recerca.

Premisses

Els actuals sistemes d'extracció automàtica de termes basats en coneixement lingüístic estan pensats per a ser usats en les llengües per a les quals són desenvolupats. Si aquests sistemes han de ser emprats per a processar llengües per a les quals no estaven pensats inicialment, cal adaptarne els patrons lingüístics i incorporar-hi eines d'anàlisi lingüística, com ara recursos d'anàlisi morfològica, etiquetatge morfosintàctic o detecció d'entitats amb nom (*name entity*) (Ananiadou, 1994a; Daille, 1997; Jacquemin, 1999).

Els sistemes d'extracció de terminologia estan preparats per a obtenir bons resultats en un àmbit d'especialitat, per a una o dues llengües, però sovint no especifiquen prou de quina manera s'han obtingut les dades, per aquest motiu són difícilment extrapolables a nous àmbits de coneixement (Estopà, 1999; Vivaldi, 2001).

Hi ha una manca de mètodes d'extracció automàtica de termes que siguin fàcilment integrables a eines de traducció assistida, traducció automàtica o extracció de terminologia i que puguin ser usats pels professionals d'aquests àmbits.

Convé prendre consciència que el resultat que s'obté fent servir mètodes d'extracció automàtica de terminologia, siguin de base lingüística, estadística o híbrida, conté candidats que no són pròpiament terminològics. Per aquest motiu, cal revisar la llista de candidats a terme d'una manera exhaustiva a fi de seleccionar les unitats que siguin representatives de l'àmbit de coneixement del qual han estat extretes; en canvi, quan el buidatge es duu a terme manualment, hi ha la feina prèvia de selecció manual dels termes, però el resultat que se n'obté són els termes propis de l'àmbit d'especialitat.

La majoria de sistemes d'extracció de terminologia no aprofiten el resultat obtingut de l'extracció prèvia per a millorar la següent extracció de can-

didats a terme. Cada vegada que s'executa el programa s'extreuen nous resultats. Fastr i Ana són dos sistemes que apliquen el mètode recursiu: a partir d'un conjunt de termes reconeguts, se n'extreuen de nous. Amb tot, es fan servir termes no validats prèviament, motiu pel qual generen resultats no adequats (Estopà, 1999, p. 56).

En el present treball hem tingut en compte les premisses exposades i proposem una combinació d'estratègies estadístiques per a ser implementades en un mètode d'extracció automàtica de termes:

- Filtratge de candidats a terme basat en *tokens* terminològics.
- Implementació de mesures d'associació lèxica aplicades a la tasca de validació manual de candidats a terme.

Els mètodes d'extracció automàtica de terminologia són diversos, i en el present treball ens centrem en l'estudi d'una combinació d'estratègies estadístiques aplicades a l'extracció de termes, essent conscients que la nostra proposta té capacitat per a ser combinada amb mètodes de caràcter lingüístic o híbrid.

Objectius

La nostra recerca se centra en dos objectius centrals:

1. Dissenyar un mètode recursiu no supervisat d'extracció automàtica de candidats a terme d'un corpus especialitzat basat en *tokens* terminològics que permeti millorar l'extracció de candidats basada en freqüència i que contribueixi en la recerca de mètodes estadístics aplicats a l'extracció automàtica de terminologia.
2. Implementar mesures d'associació lèxica aplicades a la tasca de validació manual de candidats per part d'un especialista a fi de poder millorar el nombre i la qualitat de termes extrets automàticament.

1.2 Hipòtesis

Les hipòtesis que formulem i que descrivim a continuació tenen com a objectiu comprovar quin rendiment ofereix la combinació d'estratègies estadístiques que explorem experimentalment en el present treball i que apliquem a l'extracció automàtica de termes.

D'acord amb les premisses inicials, les hipòtesis que plantegem estan relacionades amb el procés d'extracció automàtic de termes i amb la tasca de validació de candidats a terme.

Primera hipòtesi

La primera hipòtesi està relacionada amb la rendibilitat que pot oferir una extracció recursiva de termes basada en *tokens* terminològics i la formulem de la manera següent:

Un mètode recursiu d'extracció automàtica de termes basat en estratègies estadístiques permet recuperar un major nombre de termes que un mètode no recursiu basat en la freqüència.

Volem comprovar si la implementació d'un mètode recursiu d'extracció automàtica de terminologia permet millorar la identificació de termes d'un corpus especialitzat respecte l'extracció de candidats a terme basada en la freqüència.

En treballs previs l'estratègia de la recursivitat ha estat aplicada amb bons resultats en sistemes d'extracció automàtica de termes basats en mètodes lingüístics (Evans i Zhai, 1996) i mètodes híbrids (lingüístics i estadístics) (Heid, 2006) d'extracció de termes. En aquest treball analitzem la rendibilitat que ofereix l'extracció recursiva de candidats basada en *tokens* terminològics.

Segona hipòtesi

La segona hipòtesi està relacionada amb la rendibilitat que poden oferir les mesures d'associació lèxica aplicades a la tasca de validació manual de candidats a terme.

Les mesures d'associació lèxica permeten situar els termes extrets d'un corpus en les posicions inicials d'una llista de candidats a terme i, en conseqüència, redueixen el nombre de candidats que ha de ser validat manualment al final d'un procés d'extracció de termes.

Volem constatar si l'ús de mesures d'associació lèxica permet minimitzar la tasca de validació de candidats a terme, en temps i cost, que han de dur a terme els especialistes d'un àmbit d'especialitat.

Les mesures d'associació lèxica han estat aplicades en diverses ocasions a la tasca d'extracció automàtica de terminologia (Schmidt, 2001; Pazienza *et al.*, 2005) i en la detecció de col·locacions (Thanopoulos *et al.*, 2002; Evert i Krenn, 2005; Pecina i Schlesinger, 2006; Pecina, 2010; Lyse i Andersen, 2012); ara bé, no s'ha fet una anàlisi exhaustiva de l'aportació que poden fer en ser aplicades a la tasca de validació final de candidats a terme.

1.3 Metodologia

En la present recerca emprem una metodologia quantitativa per a donar resposta als objectius i les hipòtesis plantejades. Concretament, processem corpus especialitzats de l'àmbit econòmic, mèdic i de serveis socials en català, espanyol, anglès i francès per a poder contrastar el nombre de termes obtinguts amb el mètode d'extracció de terminologia proposat. Igualment, per tal d'avaluar la rendibilitat d'ús de mesures estadístiques aplicades a la tasca de validació manual de candidats, implementem onze mesures i processem els corpus esmentats per a establir relacions i regularitats entre els resultats obtinguts.

El mètode d'extracció de terminologia que proposem en aquest treball es basa en l'ús de *tokens* terminològics presents en els termes propis d'un corpus especialitzat. L'ús d'aquest tipus de *tokens* permet identificar els candidats que tenen un major caràcter terminològic. La selecció de candidats es duu a terme de manera recursiva per tal d'extreure els màxim nombre candidats a terme que estiguin constituïts per *tokens* terminològics. D'aquesta manera, s'aconsegueix reduir el nombre de resultats obtinguts al final d'un procés d'extracció automàtica de terminologia i també augmenta la probabilitat que els candidats filtrats siguin terminològics. L'aplicació del mètode en diversos corpus especialitzats i llengües variades permet constatar la rendibilitat dels resultats obtinguts comparant-ho amb el mètode de freqüència, aplicat àmpliament en processos d'extracció de terminologia.

El procés d'extracció recursiu de candidats a terme se centra en les unitats bigrams, per ser el nucli d'unitats que tendeixen a concentrar un major nombre de termes. Per unitat bigram considerem una seqüència contínua de dues paraules o *tokens* present en un corpus. Concretament, en la nostra proposta experimental considerem solament els bigrams que són de classe oberta, els quals inclouen substantius, adjectius i verbs, i no tenim en compte els bigrams de classe tancada, que inclouen conjuncions, determinants, preposicions, pronoms, verbs auxiliars i adverbis. Per aquest motiu, en la proposta que hem dissenyat no identifiquem bigrams que continguin preposicions en el seu interior.

Considerant la necessitat de millora del procés de validació manual dels candidats a terme, presentem la implementació d'onze mesures d'associació lèxica al final del procés d'extracció automàtica de termes. L'objectiu d'introduir mesures estadístiques és donar suport a l'especialista en la tasca de validar manualment els candidats. L'aplicació de mesures d'associació lèxica en l'àmbit de l'extracció automàtica de terminologia ha estat i continua essent una línia de treball molt activa i el seu ús és combinat amb mètodes lingüístics o híbrids d'extracció de termes.

1.4 Contribució de la tesi

Les principals contribucions d'aquesta tesi se centren entorn dels objectius que hem descrit més amunt.

- Aportació d'un mètode recursiu d'extracció automàtica de candidats a termes basat en *tokens* terminològics, el qual ofereix un resultat d'extracció centrat exclusivament en el caràcter terminològic que tenen els candidats extrets d'un corpus especialitzat. Aquest mètode empra com a base de filtratge els termes propis de l'àmbit d'especialitat del qual es vol dur a terme el buidatge terminològic. Aquesta aportació és especialment útil per a processar corpus provinents de llengües que compten amb pocs recursos lingüístics.

- Presenta l'ús de mesures d'associació lèxica aplicades a la tasca de validació manual de candidats a terme extrets automàticament d'un corpus especialitzat. La implementació de mesures permet rendibilitzar els resultats obtinguts en un procés d'extracció automàtica de termes, ja que es prioritzen els candidats que tenen més probabilitat de ser terminològics, els quals són endreçats a partir d'aquest rang probabilístic.

- Estudi comparatiu del rendiment que ofereixen onze mesures d'associació lèxica aplicades a la tasca de validació de candidats a terme. El resultat d'aquesta anàlisi pot servir de base per a futurs estudis que tinguin com a objecte d'anàlisi les mesures d'associació lèxica.

- Optimització dels resultats obtinguts per mitjà del filtratge recursiu de candidats a terme, fet que permet reduir el nombre de candidats que ha de ser validat manualment al final del procés d'extracció

automàtica i també disposar de candidats que tenen una major probabilitat de ser terminològics.

Els estudis preliminars i els resultats de la recerca presentada en aquest treball han estat recollits en les publicacions següents:

Oliver, T.; Vàzquez, M. (2007). *A Free Terminology Extraction Suite*. *Translating and the Computer Conference 29*. ASLIB (Association for Information Management). Londres.

Vàzquez, M.; Oliver, A. (2010). *Ús d'estratègies estadístiques per a l'extracció automàtica d'unitats terminològiques*. En J. Martí, M. Salse (coord.). *La terminologia i la documentació: relacions i sinergies*. Barcelona: Societat Catalana de Terminologia, Institut d'Estudis Catalans, p. 75-83.

Vàzquez, M.; Oliver, A. (2011). Utilisation de stratégies statistiques pour la récupération automatique de termes. En *Passeurs de mots, passeurs d'espoir. Lexicologie, terminologie et traduction face au défi de la diversité*. París: Editions des Archives contemporaines, p. 419-428.

Vàzquez, M.; Oliver, A. (2013). *Improving Term Candidate Validation Using Ranking Metrics*. *Global Journal on Technology*, p. 1348-1359.

1.5 Organització del treball

El treball està estructurat en cinc capítols, a banda d'aquesta introducció, les conclusions i el treball futur. En el capítol 2 fem una àmplia revisió dels diferents models teòrics sobre els quals es fonamenta la terminologia com a disciplina. Aquest capítol estableix la base teòrica de la proposta experimental que presentem en aquest treball.

El capítol 3 analitza en detall l'ús de mètodes lingüístics, estadístics i híbrids aplicats a l'extracció automàtica de terminologia. Així mateix, descriu quins són els àmbits en els quals s'aplica l'extracció automàtica de termes i explica quines són les limitacions que presenten les eines actuals aplicades a aquesta tasca.

Una primera aproximació experimental a l'extracció automàtica de terminologia és presentada en el capítol 4 del treball. En aquest capítol s'explica en detall la proposta d'extreure candidats a terme d'un corpus especialitzat fent ús dels *tokens* terminològics presents en els termes propis dels corpus que processem i aplicant una estratègia recursiva de filtratge. D'aquesta manera, s'obtenen candidats que tenen una alta probabilitat de ser termes i, alhora, es redueix el nombre de candidats que ha de ser revisat manualment. La recursivitat del procés permet seleccionar nous termes del corpus a mesura que s'incrementa la llista de termes de referència.

En el capítol 5 es duu a terme una descripció detallada de les onze mesures d'associació lèxica que són implementades de manera experimental en aquest treball. En primer lloc fem una descripció general de la mesura i aportem la corresponent fórmula de càlcul. Seguidament donem informació relacionada amb l'ús de cada mesura estadística en l'àmbit del processament del llenguatge natural.

L'aplicació experimental de les mesures d'associació lèxica és duta a terme en el capítol 6, el qual descriu la implementació de mesures aplicades a la tasca de validació de termes. Concretament, la proposta experimental es basa a processar els resultats obtinguts amb el mètode d'extracció automàtica de termes descrits en el capítol 4 amb les onze mesures d'associació lèxica. La base d'aquesta proposta se centra a identificar quins són els candidats que tenen una major probabilitat de ser termes aplicant-hi els criteris d'importància i força d'associació, els criteris basats en la teoria de la informació i els criteris heurístics que tenen associats cada una de les mesures. Així, els candidats són ponderats per un nivell de rang segons cada mesura, fet que permet poder destriar quin candidat té

major probabilitat de ser terminològic.

El capítol 7 descriu les conclusions que es desprenen de les propostes experimentals plantejades.

I el capítol 8 presenta la continuació de la recerca descrita en el present treball.

Capítol 2

FONAMENTS TEÒRICS DE LA TERMINOLOGIA

Els orígens de la terminologia es remunten al segle XVIII, moment en el qual els científics van mostrar la seva preocupació per poder regular la creació de denominacions que garantissin una comunicació unívoca. Aquest interès se centrà en terrenys com ara la química, de la mà de Lavoisier (1793) i Berthollet (1803), o de la botànica i la zoologia, amb Linné (1736). El treball de Linné, *Fundamenta botanica* (1736), és considerat la primera terminologia sobre botànica sistemàtica i coordinada.

Al segle XX, amb l'arribada de la revolució tecnològica i industrial, s'estableixen les bases lingüístiques de la terminologia com a matèria autònoma i interdisciplinari, gràcies a l'aportació de l'escola de Viena, la de Praga i la soviètica. A partir del treball que fan aquestes escoles, la terminologia s'estén tant pel que fa als postulats teòrics com al vessant pràctic. Durant aquest període hi ha els treballs i les iniciatives portats a terme en centres de recerca del Canadà (Quebec), França, els països nòrdics i la Gran Bretanya (a l'UMIST, l'Institut de Ciència i Tecnologia de la Universitat de Manchester, University of Manchester Institute of Science and Technology).

2.1 Sobre terminologia

Terminologia és una denominació que presenta ambigüitat semàntica i compta amb diverses accepcions. Segons el context en el qual es faci servir, Rondeau (1984) i Felber (1984) indiquen que pot referir-se...

> a) al conjunt de termes que representen els conceptes d'un àmbit d'especialitat.
>
> b) als mètodes de recollida i classificació dels termes, de creació neològica, de normalització, de difusió, o a la publicació en la qual els conceptes d'un àmbit d'especialitat són representats per termes.
>
> c) a una ciència que té per objecte l'ordre lingüístic, interdisciplinària, relacionada amb els conceptes i les seves representacions i que participa de la lingüística, la lògica, l'ontologia i la informàtica.

Per a Sager, terminologia és l'estudi i l'àmbit d'activitat relacionats amb l'emmagatzematge, la descripció, el processament i la difusió de termes, i en distingeix tres sentits (Sager, 1990, p. 3):

> 1. The set of practices and methods used for the collection, description and presentation of terms.
>
> 2. The set of premises, arguments and conclusions required for explaining the relationships between concepts and terms which are fundamental for a coherent activity under 1.
>
> 3. A vocabulary of a special subject field.

La denominació de terminologia pot ser usada per a descriure mètodes de recopilació, difusió i estandardització de termes; com a teoria, i és quan la paraula adquireix el seu significat com a resultat de l'aproximació feta per Wüster (1979a), i per a descriure el vocabulari d'un àmbit d'especialitat. En el present treball ens centrem en el conjunt de termes que representen els conceptes d'un àmbit d'especialitat, segons Rondeau, Felber i Sager.

El treball terminològic té com a base l'anàlisi, la definició dels conceptes i la manera d'anomenar-los, i la publicació dels resultats d'aquest treball és la base de la terminografia. El treball terminològic està marcat per les normes metodològiques orientades a la tasca de recopilació i emmagatzematge de terminologia, normes que estableixen principis aplicables al treball terminològic amb la finalitat que els terminòlegs elaborin els seus repertoris i bases de dades a partir dels mateixos principis per a obtenir millors resultats i que aquests puguin ser comparables, de manera que permeti l'intercanvi de dades. Concretament, el Comitè Tècnic 37 de la International Organization for Standardisation (ISO) (ISO, 2012) té la missió d'estandarditzar principis, mètodes i aplicacions relacionades amb la terminologia, les indústries de la llengua i els diferents recursos terminològics. Aquest comitè tècnic ha preparat normes específiques amb l'objectiu d'establir un estàndard comú en relació amb el treball terminològic:

- ISO 704:2009 Terminology work. Principles and methods
- ISO 860:2007 Terminology work. Harmonization of concepts and terms
- ISO 1087-1:2000 Terminology work. Vocabulary. Part 1: Theory and application

La norma ISO 1087-1:2000, i la corresponent norma UNE-ISO 1087-1:2009 (UNE, 2009) traduïda al català pel TERMCAT, Centre de Terminologia, defineix terminologia de la manera següent:

> a) Ciència que estudia l'estructura, la formació, el desenvolupament, l'ús i la gestió de la terminologia en diverses àrees temàtiques.
>
> b) Conjunt de denominacions que pertanyen a una llengua d'especialitat.

Amb la incorporació de la informàtica en els processos de gestió terminològica, han aparegut conceptes per a denominar la relació que s'estableix entre la tasca pròpiament terminològica i les noves possibilitats de processament de corpus que ofereixen les tecnologies lingüístiques. Així, s'utilitza el concepte *terminòtica* per a definir la relació que hi ha entre la informàtica i la terminologia, relació que permet crear eines de tractament i emmagatzematge de dades terminològiques (Cabré, 1992). En l'àmbit de la terminòtica, Depecker (1998) fa servir el concepte *termatique*, que defineix com "la discipline spécifique à l'application de l'informatique au terme". D'aquesta manera, *terminòtica* i *termatique* són dos conceptes que pertanyen a la gran àrea (o ampli àmbit d'especialitat) de les indústries de la llengua, és a dir, un conjunt d'activitats encaminades a concebre, fabricar i comercialitzar maquinari i programari que permet d'interpretar i de processar el llenguatge natural, escrit o parlat. La recerca en aquest àmbit se centra en la preparació de programes que permeten la verificació i la correcció ortogràfica i gramatical, la gestió de bancs de dades textuals i de coneixement terminològic (Conceição, 2005).

En aquesta mateixa línia sorgeix el concepte *terminologia computacional*, que s'utilitza per a identificar la recerca centrada en la millora de les estratègies lingüístiques i estadístiques d'extracció automàtica d'unitats terminològiques presents en un text perquè els resultats puguin ser aprofitats per altres aplicacions, per a reagrupar variants de candidats per tal d'oferir una llista acurada de resultats i per a identificar la tasca de cercar informació semàntica i conceptual dels termes o representar relacions conceptuals entre els termes (Bourigault *et al.*, 2001, p. IX). Així mateix, aquest concepte és definit per Foo de la manera següent:

> "The field of *Computational Terminology* (CT) studies how computational methods can be of use when performing Terminology Work." (Foo, 2012, p. 2)

Aquestes denominacions consoliden el paper destacat que té la tecnologia en el processament de la terminologia present en els corpus especialitzats.

Concepte i terme

Un concepte és una construcció mental que serveix per a classificar els objectes individuals del món exterior o interior a través d'un procés d'abstracció més o menys arbitrari (Aguilar, 2001). Bessé considera que un concepte és una unitat abstracta que té les característiques d'objectes concrets o abstractes, i ho defineix de la manera següent:

> "An abstract unit which consists of the characteristics of a number of concrete or abstract objects which are selected according to specific scientific or conventional criteria appropiate for a domain." (Bessé *et al.*, 1997, p. 119)

En la teoria tradicional de la terminologia, la relació entre concepte i terme és formalment equivalent a la relació que s'estableix entre significat i paraula; l'estructura descriptiva dels conceptes (conjunt de característiques) és idèntica a l'estructura descriptiva simple dels significats (conjunt de trets semàntics), i el detall dels sistemes conceptuals descrits fins al moment en els estudis de terminologia no van més enllà del detall dels sistemes semàntics o conceptuals presents en els estudis que no tenen relació amb la terminologia (Kageura, 2002).

La norma ISO 1087-1:2000, i la corresponent norma UNE-ISO 1087-1:2009 (UNE, 2009), defineix *concepte* com a "unitat de coneixement formada per una combinació única de característiques", entenent *característica* en el sentit de "propietat abstracta d'un objecte o d'un conjunt d'objectes".

La principal característica d'un terme és ser monoreferencial en sentit ampli, ja que un terme normalment fa referència a un concepte específic d'un àmbit d'especialitat determinat. El coneixement d'un àmbit d'especialitat és transmès de manera natural als diferents subllenguatges (llenguatge per a finalitats específiques o per a àmbits lingüístics restringits) (Hoffmann, 1998), fet que mostra restriccions en els nivells lèxic, sintàctic i semàntic. Cada subllenguatge té els seus propis conceptes, i els terminòlegs els

estructuren en un sistema organitzat tenint en compte les relacions que mantenen els conceptes (relació genèrica, de causa, etc.). En aquest sentit, els termes són la realització lingüística dels conceptes en una determinada comunicació i són organitzats en sistemes de termes que idealment reflecteixen el sistema conceptual associat.

En parlar de terme, cal considerar que no hi ha una única definició, i això és degut a la varietat d'enfocaments que existeix segons l'objecte d'estudi. Els terminòlegs tradicionals consideren que la noció de *terme* pot fer referència als ítems lèxics que tenen una referència especial en un àmbit d'especialitat específic (Sager, Juan C. *et al.*, 1980); pot ser l'equivalent al signe lingüístic de Saussure (combinació de significant i significat) (Rondeau, 1984, p. 19); pot ser una etiqueta o un símbol lingüístic per a un concepte (Felber, 1983, p. 8); pot descriure la combinació entre *denominació* i *noció*, tenint en compte la distinció que Rondeau (1984) fa entre *denominació* (etiqueta) i *noció* (concepte), o pot referir-se exclusivament a *etiqueta* (Wüster, 1979a). Aquestes distincions s'han fet entre termes tècnics que són usats en un determinat àmbit d'especialitat i termes generals que s'usen en més d'un àmbit. I les distincions també s'han fet entre termes el significat dels quals ha estat consensuat i paraules el significat de les quals no ha estat consensuat. La tendència dels terminòlegs tradicionals és establir una correlació d'un a un entre concepte i terme amb l'objectiu de reduir l'ambigüitat i millorar la comunicació, crear jerarquies conceptuals per a representar el coneixement, fer classificacions –compilació de terminologies estandarditzades. Ara bé, a la pràctica és complex d'aplicar aquestes distincions per a decidir si una unitat lèxica en un text és usada com a paraula de la llengua general o com un terme. Sembla que únicament una persona experta en un àmbit d'especialitat pot identificar els termes que són propis d'aquest àmbit. I amb l'augment de la interdisciplinarietat, si dos àmbits d'especialitat tenen termes en comú, aquests termes, han de ser considerats simplement paraules, com diu Sager?

Un terme també és considerat com la intersecció entre la part conceptual (contingut semàntic) i la part lingüística (Lauriston, 1996); la representació lingüística dels conceptes en un determinat àmbit d'especialitat (Frantzi i Ananiadou, 1997), o com una unitat lèxica formada per una paraula o més d'una que representa un concepte dins un àmbit d'especialitat, tal com defineix Bessé:

> "A lexical unit consisting of one or more than one word which represents a concept inside a domain. Note: There are (a) general terms of a subject field which are used in general descriptions, instructions and textbooks, patent descriptions, and other non-industry specific terms. These terms usually have a long life in which they undergo changes of meaning in the form of extensions or narrowing of extension. For example, "lamps" embrace the oil lamps of the ancients Greeks, gas lamps, electric lights, etc. (b) Craft- or industry-specific and even firm-specific terms which are more specialised. Many of these are homonyms requiring different definitions according to the subject field of their application. (c) Product-specific terms, frequently designations of material entities, which have a limited life because they are firmly linked to a manufactured physical object which may be superseded by a similar object, which, however, may be given a different designation in order to stress its difference." (Bessé *et al.*, 1997)

I en la norma ISO 704:2009 Terminology work. Principles and methods hi ha recollida la definició de terme següent:

> "Designation consisting of one or more words representing a general concept in a special language in a specific subject field. A simple term contains only one root, while a term containing two or more roots is called a complex term. [...] Ideally, when precise and accurate communication is required in a given special language, especially in fields of science and technology, the objective of term—concept assignment is to ensure that, within a given context, a given term is attributed to one concept (monosemy) and that a gi-

> ven concept is represented preferably by one preferred term. This condition reduces ambiguity while homonymy and synonymy can lead to ambiguity." (ISO, 2009)

En aquesta norma s'especifica que quan s'ha de portar a terme una comunicació precisa en un àmbit d'especialitat, especialment en ciència i tecnologia, l'objectiu d'assignar termes a conceptes serveix per a assegurar que, en un determinat àmbit d'especialitat, un terme és atribuït a un concepte (monosèmia) i que un concepte és representat preferentment per un terme. Malgrat aquesta condició redueix l'ambigüitat, cal considerar que l'homonímia i la sinonímia també poden introduir-ne.

Des del punt de vista dels especialistes, la consideració de què és un terme també varia. Segons els terminòlegs, les paraules esdevenen o adquireixen el caràcter de terme en ser usades en àmbits d'especialitat. Segons els especialistes en processament del llenguatge natural, el lèxic i la gramàtica usats en àmbits d'especialitat són restringits. Dues maneres de considerar què és un terme, malgrat compartir una mateixa visió: els termes són diferents de les paraules. Essent aquesta la situació, cal posar el punt de mira en quins són els termes adequats en cada context, ja que és molt difícil establir una diferència entre terme i paraula sense tenir en compte el context en el qual seran usats.

En la diversitat de consideracions en relació a la definició de terme, es mostra una posició consensuada en el fet de considerar que un terme és una unitat que s'esdevé en un àmbit d'especialitat. Sobre el tipus d'unitat, hi ha una posició que considera que és una combinació entre denominació i noció (significant i significat o part conceptual i part lingüística) i n'hi ha una altra que considera que és una etiqueta (símbol lingüístic).

RESUM DE LES CONSIDERACIONS SOBRE TERME

Sager (1980). Ítem lèxic que té una referència especial en un àmbit d'especialitat específic.

Rondeau (1981). L'equivalent al signe lingüístic de Saussure (combinació de significant i significat).

Felber (1983). Etiqueta o un símbol lingüístic per a un concepte.

Rondeau (1984). Combinació entre *denominació* (etiqueta) i *noció* (concepte).

Wüster (1985). Fa referència exclusivament a *etiqueta*.

Lauriston (1996). Intersecció entre la part conceptual (contingut semàntic) i la part lingüística.

Frantzi (1997). Representació lingüística dels conceptes en un determinat àmbit d'especialitat.

Bessé (1997). Unitat lèxica formada per una paraula o més d'una que representa un concepte en un àmbit d'especialitat.

Terme i paraula

A l'hora d'establir una frontera entre terme i paraula, també es constata que hi ha una gran diversitat de posicions. Ja Rondeau (1984) reivindicava que hi ha diferència entre les paraules i els termes; ara bé, a banda de declarar que els termes són usats en àmbits de coneixement específics, no ofereix cap més distinció.

Per la seva banda, Felber defineix tres tipus de símbols lingüístics:

> "The word. [...] The *word* can have a multiplicity of nondefined meanings and shades of meanings or can be used for naming ob-

> jects. The concrete meaning of a word is given by the context; in other words, it is dependent on context. [...]
>
> The term. [...] The *term* is a linguistic symbol assigned to one or more concepts (defined meanings). The meaning of a *term* which is the concept, is dependent on the position of this concept in the system of concepts concerned. [...]
>
> The thesaurus word. [...] is a word, for the most part a *term* or a name, that is used for indexing and retrieval of information in information systems." (Felber, 1983, p. 8)

Felber diu que les paraules tenen significat solament a partir del context en què es troben, però no concreta quines paraules poden ser descrites com a símbols lingüístics. Així mateix, cal tenir present que hi ha moltes paraules que no poden ser considerades terme, però que tenen un significat definit, per exemple, els colors, els conceptes abstractes (felicitat, tristesa) o els objectes concrets (cadira o conill). I és poc probable que Felber considerés aquestes paraules com a termes, perquè llavors qualsevol paraula que tingui un significat definit seria un terme. Quan defineix *terme* diu que aquests són diferents de les paraules perquè tenen un significat definit, però amb aquest criteri no n'hi ha prou per a fer la distinció perquè hi ha paraules que també tenen un significat definit. I, finalment, la distinció entre paraules, termes i tesaurus és poc aclaridora de quina diferència hi ha entre aquestes tres categories (Pearson, 1998).

Sager estableix la distinció entre terme i paraula des d'una dimensió cognitiva de la manera següent:

> "The lexicon of a special subject language reflects the organisational characteristics of the discipline by tending to provide as many lexical units as there are concepts conventionally established in the subspace and by restricting the reference of each such lexical unit to a well-defined region. Besides containing a large number of items which are endowed with the property of special reference the lexicon of a special language also contains items of general reference

> which do not usually seem to be specific to any discipline or disciplines and whose referential properties are uniformly vague or generalised. The items which are characterised by special reference within a discipline are the 'terms' of that discipline, and collectively they form its 'terminology'; those which function in general reference over a variety of sublanguages are simply called 'words' and their totality the 'vocabulary'." (Sager, 1990, p. 19)

En la definició que aporta Sager s'observen alguns dels problemes que hi ha a l'hora de distingir entre paraula i terme. Concretament, en la primera frase es confirma el que també diu Wüster, que el lèxic reflecteix l'estructura conceptual d'un àmbit temàtic, que la referència a cada unitat lèxica és restringida a l'àmbit en qüestió i que els conceptes són convinguts (establerts de manera convencional). I continua dient que el lexicó d'una determinada llengua té dues classes d'ítems: els ítems amb referències especials, que són els termes, i els ítems amb referències generals, que són les paraules. Aquesta darrera classe consta d'ítems que no són habitualment "specific to any discipline or disciplines" i les seves propietats referencials són "uniformly vague or generalised". I acaba dient que han de ser classificats com a paraules. No dóna cap exemple d'ítems de referència general, però es pot considerar que pertanyen al lexicó d'un determinat àmbit temàtic. Sager fa servir *paraula* com a categoria general per a tots els ítems lèxics que no encaixen amb la seva classificació de terme (Pearson, 1998).

Des d'una dimensió lingüística, Sager considera que "Terms are the linguistic representation of concepts" (Sager, 1990, p. 57). Sager (1998), a l'hora d'establir la frontera entre paraules i termes, també constata que "it can happen that non-specialists consider a word to be a term which is, however, only a general word for the specialist; equally, it can happen that specialists use terms which their non-specialist audience take to be words in the general language". A més, "the possibility of many lexical units to function both as words and as terms may even be a question of individual choice and interpretation of the speaker and listener". Els termes constantment interactuen amb les paraules de la llengua general

perquè comparteixen les mateixes formes lingüístiques. En aquest sentit, s'hi afegeix el criteri específic d'especialistes i no especialistes per a determinar el caràcter de terme o paraula en un àmbit d'especialitat concret.

Recentment hi ha lingüistes que centren les diferències entre terme i paraula en la significació del terme i en el que representen les nocions de caràcter conceptual, de significat, de concepte, de sentit o de definició. Concretament, Slodzian (2000) comenta que l'accés continuat als textos especialitzats en certa manera els ha "desespecialitzat" i ha posat en evidència fenòmens difícilment compatibles amb una visió clàssica del terme: d'una banda, la variació terminològica en els textos i, d'altra banda, l'augment de la polisèmia a mesura que els àmbits temàtics es barregen, com es constata en els treballs multidisciplinaris. Així, es pot dir que el coneixement específic d'un àmbit es troba tant en els textos escrits per la comunitat científica o tècnica com en el terme creat per un autor i les definicions que aquest ha creat. Slodzian fa referència a l'aproximació textual per a poder descriure i estudiar el funcionament real de les unitats en el discurs. Aquesta autora crea el concepte *candidat a terme* i defensa que el paper de l'especialista consisteix a fer una selecció final dels termes que s'han de tenir en compte de llistes establertes pels terminòlegs a partir de textos reals. En aquest context, doncs, la frontera entre paraula i terme es desdibuixa. Slodzian també indica l'obsessió que hi ha des del punt de vista de la terminologia clàssica de distingir terme de paraula i que aquesta diferència té com a element clau la significació.

Així mateix, Cabré (2000) considera que termes i paraules formen part d'una mateixa unitat: la unitat lèxica. I ho considera així tenint en compte que la terminologia representa el conjunt d'unitats utilitzades en la comunicació especialitzada; que els termes no són unitats autònomes que constitueixen un lèxic especialitzat separat d'un lèxic general, sinó que són un conjunt de trets de significació associats a les unitats lèxiques, les quals activen el caràcter de paraula o terme segons les característiques pragmàtiques de la situació en què s'utilitzen; que l'activació del caràcter de terme o paraula d'una unitat lèxica s'obté per mitjà de la selecció de

conjunts de trets; que els termes reals són potencialment polisèmics, ja que llur significat es pot ampliar i multiplicar en diferents àmbits d'especialitat, i que el valor d'un terme està determinat per la seva aparició en un àmbit d'especialitat.

Sager (2000) comenta que els termes són unitats lèxiques que es presenten en forma de nom i que estan associats a una significació i a una referència més precises que les de les paraules, ja que la seva funció és la de designar els conceptes clarament identificats en un determinat àmbit temàtic. Els termes es distingeixen de les paraules pel fet que són escollits i formats expressament per a designar conceptes que els parlants han decidit distingir, perquè volen que la referència sigui més precisa que la que s'obté amb les paraules.

I des del punt de vista de la comunicació, Costa (2006) comenta que el terme es comporta a nivell sintàctic, morfològic i lèxic, a la llengua i al discurs, com qualsevol unitat lèxica. Aquesta mateixa idea també és recollida per Kocourek:

> "Les termes et les non-termes du fond lexical entier d'une langue ont en commun l'appartenance à cette langue, d'où résulte leur dépendance au système phonologique, morphologique et lexical de cette langue, leur subordination, dans le contexte, aux lois syntaxiques et aux tendances sémantiques et, enfin, leur validité inégale dans le temps, dans l'espace et la réalité sociale." (Kocourek, 2001, p. 296)

En resum, la idea de Wüster i els seus deixebles de mantenir oposades les nocions de terme i paraula pel fet que la terminologia havia de ser diferenciada de la lexicologia i perquè el terme era considerat com una eina de comunicació, idealment unívoc i monoreferencial el sentit del qual era establert respecte a un àmbit d'especialitat, es considera limitada. Actualment amb aportacions com les de Cabré (1992, 1999a), Gambier (1993), Slodzian (1993, 1995), Condamines (1994, 1995), Delavigne Delavigne (2001), Gaudin (2003) i Gaudin i Alexandru (2005) es presenta un nou

enfocament en el qual el context discursiu pren un paper decisiu. Així, en admetre la polisèmia del terme i la necessitat del context per a la desambiguació del sentit, es plantegen noves formes de determinar la referència i nous criteris per a identificar les paraules el sentit especialitzat de les quals s'activa pel context i en el context.

RESUM DE LA DISTINCIÓ ENTRE TERME I PARAULA

Felber (1983). Les paraules depenen del context. Els termes són símbols lingüístics assignats a un concepte o a més d'un.

Rondeau (1984). Els termes són usats en àmbits de coneixement específics.

Sager (1990). Els elements que són caracteritzats per una referència especial en una disciplina són termes. Els elements que tenen una referència general en diversos subllenguatges són paraules.

Sager (1998). Una persona no-especialista pot considerar terme una paraula que és general per a un especialista, i a la inversa.

Slodzian (2000). La diferència entre terme i paraula rau en la significació.

Cabré (2000). Termes i paraules formen part d'una mateixa unitat: la unitat lèxica.

Sager (2000). Els termes es distingeixen de les paraules pel fet que designen conceptes que els parlants han decidit distingir, perquè volen que la referència sigui més precisa que la de les paraules.

Kocourek (2001), Costa (2005). El terme es comporta a nivell sintàctic, morfològic i lèxic, a la llengua i al discurs, com qualsevol unitat lèxica.

2.2 Teories entorn de la terminologia

Seguint el plantejament fet per Auger (1988), els grans períodes de la terminologia es poden classificar de la manera següent:

- Els orígens (1930-1960)
- L'estructuració (1960-1975)
- L'expansió (1975-1985)
- L'ampliació (1985-en endavant)

El període inicial d'estudi de la terminologia (1930-1960) es caracteritza pel disseny de mètodes orientats a la formació sistemàtica de termes i és el moment en què neix la terminologia moderna, de la mà de Wüster. En la seva tesi doctoral, Wüster (1931) presenta arguments per a sistematitzar els mètodes de treball en terminologia, estableix diferents principis per a treballar amb termes i esbossa els punts principals de la metodologia per a processar les dades terminològiques.

En el segon estadi de desenvolupament (1960-1975), les innovacions en el terreny de la terminologia es produeixen amb l'aparició dels ordinadors i les tècniques documentals. En aquest moment apareixen les primeres bases de dades i comencen a establir-se els principis del processament de la terminologia. És un període en el qual els esforços se centren en l'estandardització de la terminologia. En aquest període Wüster publica el seu diccionari, *The Machine Tool* (1968) (Wüster, 1968), obra que suscità el seu interès amb vista a fer una reflexió teòrica entorn de la terminologia.

La tercera època (1975-1985) està marcada per la planificació lingüística i els projectes terminològics. El paper que juga la terminologia en la modernització d'una llengua s'esdevé en aquest període. I l'explosió dels ordinadors fa possible una nova manera de processar les dades terminològiques. En aquest moment es publica l'obra pòstuma de Wüster.

En el darrer període (1985-en endavant) la informàtica és una de les principals causes de canvi en la terminologia. Els terminòlegs tenen al seu abast eines i recursos que s'adapten a les seves necessitats. En aquest moment apareix un nou mercat, el de les indústries de la llengua, en el qual la terminologia juga un paper principal. Es consolida la cooperació internacional, fet que permet l'intercanvi d'informació i la formació de terminòlegs per mitjà d'aquesta cooperació, i també la planificació lingüística. Aquest darrer període es caracteritza per les diferents aproximacions de tendència social, comunicativa i cognitiva –com ara la socioterminologia, la teoria comunicativa de la terminologia o l'aproximació sociocognitiva– i les aplicacions de la recerca terminològica que beneficien les eines informàtiques i les indústries de la llengua.

Aquests quatre grans períodes de la terminologia esdevenen en paral·lel a diferents enfocaments teòrics, marcats per l'evolució en la manera de considerar el treball terminològic. Seguidament fem una descripció detallada dels plantejaments teòrics que sorgeixen dels anys trenta del segle passat ençà.

PLANTEJAMENTS TEÒRICS ENTORN DE LA TERMINOLOGIA

Orígens (1930-1960)
Teoria tradicional de la terminologia (Wüster)

Estructuració (1960-1975)

Expansió (1975-1985)
Socioterminologia (Gambier, Gaudin)

Ampliació (1985- en endavant)
Teoria Sociocognitiva Terminologia (Temmerman)
Teoria Comunicativa de la Terminologia (Cabré)
Terminologia Textual (Bourigault i Slodzian)

2.2.1 Teoria tradicional de la terminologia

En els anys trenta del segle passat Wüster, considerat el pare de la terminologia moderna i fundador de l'escola de Viena, ja va esbossar la metodologia de treball amb dades terminològiques amb la publicació de la seva tesi doctoral. D'entrada, la preocupació de Wüster és bàsicament metodològica i normativa, no teòrica. El seu interès per la teoria vindrà més endavant, com a fruit de reflexió del procés de treball en la confecció del seu diccionari, *The Machine Tool* (1968). En la seva obra pòstuma de 1979, *Einführung in die Allgemeine Terminologielehre und terminologische Lexikographie* (Wüster, 1979a), hi ha el compendi de la seva teoria, anomenada Teoria General de la Terminologia (TGT) en referències posteriors (Wüster, 1979b), que serà desenvolupada i radicalitzada posteriorment pels membres de l'escola de Viena.

Els motius que Wüster té per a endinsar-se en l'àmbit de la terminologia són merament pràctics, és a dir, superar els obstacles de la comunicació professional provocats per la imprecisió, diversificació i polisèmia del llenguatge natural. Wüster considera la terminologia com un instrument de treball que ha de servir de manera eficaç a la desambiguació de la comunicació científica i tècnica.

La distinció que Wüster fa dels termes i les paraules de la llengua general se centra en tres aspectes: primer, en lexicografia la unitat lèxica és el punt de partida habitual i en terminologia el nucli principal és el concepte. El concepte ha de ser considerat de manera aïllada de la seva etiqueta terme. Els conceptes existeixen independentment dels termes i també independentment de qualsevol llengua. Segon, els terminòlegs s'interessen en el vocabulari de manera aïllada, sense considerar la morfologia o la sintaxi, i això contribueix al fet que Wüster percebés els termes de manera separada de les paraules, diferents no solament pel que fa al significat, sinó també per la seva natura i ús. Tercer, els terminòlegs, d'acord amb Wüster, tenen interès a proposar normes per a l'ús del llenguatge, cosa que ha estimulat la creació de vocabularis estandarditzats. Malauradament, el

fet de tenir una terminologia estandarditzada no és cap garantia que sigui usada. La noció d'estandardització de l'ús és un punt central de la teoria de Wüster, perquè els termes estandarditzats són usats per a representar les estructures conceptuals en què se sustenten els àmbits de coneixement.

La teoria de Wüster, que defineix la terminologia com un camp de trobada entre la lingüística, la ciència cognitiva, la ciència de la informació, la comunicació i la informàtica, estableix un objecte d'anàlisi i unes funcions de treball molt restrictives, perquè limita el seu objecte de treball a les unitats unívoques normalitzades pròpies dels àmbits cientificotècnics, redueix la terminologia a la recopilació de conceptes per a la normalització dels termes, situa els àmbits especialitzats en la ciència i la tècnica, i limita els seus objectius amb la finalitat d'assegurar la univocitat en la comunicació professional, sobretot en el terreny internacional (Cabré, 2002).

Cabré resumeix els elements fonamentals de la teoria de Wüster de la manera següent:

- La terminologia es concep com una matèria autònoma i es defineix com un camp d'intersecció format per les "ciències de les coses" i per altres disciplines com la lingüística, la lògica i la informàtica.
- L'objecte d'estudi d'aquesta teoria són els conceptes, transmesos a través d'unitats de designació, unitats lingüístiques (denominatives i designatives alhora) i unitats no lingüístiques (exclusivament designatives). Aquestes unitats són específiques d'un àmbit d'especialitat i el seu ús està restringit a aquest àmbit.
- Els termes es defineixen com les denominacions lingüístiques dels conceptes, així un terme és la unitat (lingüística i no lingüística) que designa un concepte.
- Els termes s'analitzen a partir del concepte que representen; per tant, s'assumeix que el concepte precedeix la denominació.

- Els conceptes d'un mateix àmbit d'especialitat mantenen entre ells relacions de diferent tipus. El conjunt de les relacions entre els conceptes constitueix l'estructura conceptual d'una matèria. El valor d'un terme s'estableix pel lloc que ocupa en l'estructura conceptual d'una matèria.

- L'objectiu és estudiar els termes des de la perspectiva de la normalització conceptual i denominativa, monolingüe, en el cas de la comunicació professional nacional, o plurilingüe, en el cas de la comunicació internacional.

- La finalitat aplicada de la normalització terminològica és garantir la precisió i la univocitat de la comunicació professional –estrictament professional– mitjançant l'ús dels termes normalitzats.

Felber (1984), deixeble i compilador de l'obra pòstuma de Wüster, recull tres característiques específiques en relació amb la teoria de la terminologia:

- Any terminology work starts with concepts. It aims at the strict delimitation of concepts. The sphere of concepts is independent of the sphere of terms.

- Only the terms of concepts, i.e. the terminologies, are of relevance to the terminologist, not the rules of inflections and the syntax.

- The terminological view of language is a synchronic one, i.e. for terminology the present meanings of terms are important. For terminology the system of concepts is what matters in language.

L'enfocament tradicional de la teoria de la terminologia fa referència a la relació entre conceptes i termes, començant pels conceptes i centrant-se en l'estat de l'estructura conceptual i la seva representació. Així, el concepte pren un paper central i és definit com a "element del pensament", el qual "consisteix en un conjunt de característiques".

El plantejament de fons de la TGT és normalitzar i fixar la relació entre terme i concepte per a la comunicació professional internacional, per la qual cosa la motivació d'aquesta teoria de la terminologia és la planificació lingüística, amb una clara orientació vers la prescripció. Aquesta delimitació de l'objecte terminològic i la seva funció merament denominativa (dimensió representacional), fa necessari un enfocament renovador que tingui en compte la complexitat de la unitat terminològica i també la seva dimensió social i comunicativa (Gómez González-Jover, 2007).

2.2.2 Vers un nou model teòric

La consideració de la Teoria General de la Terminologia com a model de treball tradicional comença a canviar en no acabar de resoldre els problemes per als quals s'havia plantejat i no poder donar resposta a les necessitats terminològiques pel que fa a l'ús real de la llengua i les seves necessitats comunicatives. L'objectiu de construir una terminologia única i estàndard esdevé una tasca fictícia, perquè no és possible estandarditzar els usuaris ni les situacions en què es produeix la comunicació (Gómez González-Jover, 2007).

En el procés de revisar la teoria i la pràctica terminològica hi ha quatre fonts principals que hi influeixen: la sociolingüística teòrica, la sociolingüística aplicada, la lingüística general i la lingüística de corpus (Gaudin, 2003). En aquest procés de qüestionar la teoria tradicional de la terminologia, han sorgit diverses veus crítiques que, després de posar en relleu les limitacions de la teoria wüsteriana, han fet propostes innovadores. Des de 1990 autors com Sager, Desmet o Kageura han exposat les deficiències dels mètodes i principis de la terminologia tradicional i han presentat noves propostes entorn d'aspectes diversos:

- Els conceptes clau en terminologia
- Les dimensions de la terminologia
- La variació terminològica

En relació amb els *conceptes clau en terminologia*, s'han fet noves aportacions sobre el concepte, la definició, la monosèmia i la sincronia. En aquest sentit, Kageura (2002) considera que és discutible la prioritat que es dóna als conceptes per sobre dels termes. Si l'àmbit dels conceptes és independent de l'àmbit dels termes, tot pot ser situat en l'esfera dels conceptes sense tenir la certesa que els conceptes tindran una relació rellevant en l'àmbit dels termes. Així mateix, comenta que no es pot acceptar la consideració de la teoria tradicional que diu que "*only the terms of concepts are of relevance*" per a la terminologia. I també que el punt de vista terminològic de la llengua no ha de ser únicament de tipus sincrònic. Un estudi diacrònic dels termes pot ser perfectament considerat un estudi teòric de la terminologia.

Pel que fa a les *dimensions de la terminologia*, es veu la necessitat d'incorporar a les dues dimensions que s'han considerat tradicionalment en terminologia –dimensions cognitiva i lingüística– una tercera dimensió, la comunicativa, que permeti estudiar els termes en un context real i variant (Sager, 1990). Dubuc (Dubuc, 1978, p. 14), en plena època d'expansió de la TGT, expressa que la terminologia té funcions d'expressió i comunicació. I, per una altra banda, Sager (Sager, 1990, p. 13) observa tres dimensions rellevants en la terminologia: la cognitiva, la lingüística i la comunicativa:

1. A cognitive one which relates the linguistic forms to their conceptual content, i.e. the referents in the real world.

2. A linguistic one which examines the existing and potential forms of the representation of terminologies.

3. A communicative one which looks at the use of terminologies and has to justifiy the human activity of terminology compilation and processing.

Des d'un punt de vista cognitiu, la terminologia fa referència al llenguatge emprat en un determinat àmbit d'especialitat, per la qual cosa hi haurà

tants llenguatges d'especialitat com àrees de coneixement o d'activitat hi hagi en una determinada comunitat lingüística. Terminològicament parlant, el lexicó d'una llengua representa l'estructura del coneixement de cada àmbit d'especialitat o disciplina. I cada estructura del coneixement representa la relació entre conceptes. Acostar l'estudi de la terminologia a la seva dimensió cognitiva demana entendre l'estructura del coneixement per a tenir una completa i coherent representació de la natura, el comportament i la interacció dels conceptes i els termes que hi estan relacionats. Des del vessant lingüístic, els termes són la representació lingüística dels conceptes. Un concepte pot tenir tantes representacions lingüístiques com tantes situacions comunicatives hi hagi que requereixin diferents formes lingüístiques. Des d'un punt de vista comunicatiu, els termes funcionen segons el model comunicatiu que hi ha entre dos especialistes, i l'ús de la llengua afecta la natura i el comportament dels termes.

Segons Desmet (Desmet, 1995, p. 98), les principals funcions de la terminologia se centren en l'estudi de la cognició, la comunicació i la representació del coneixement especialitzat, i també en el fet de proporcionar models de representació del coneixement que puguin assegurar una correcta transferència del coneixement.

I pel que fa a la *variació terminològica*, les tendències més recents donen importància a les funcions de representació i de transmissió de coneixement de la terminologia. Així, segons Cabré (1995), la representació del coneixement està en relació amb la documentació, l'enginyeria de la llengua i la lingüística informàtica, i la transmissió del coneixement està en relació amb les necessitats de comunicació, la mediació comunicativa, la planificació, les polítiques lingüístiques i culturals. La funció de representació implica tenir en compte la variació terminològica i la variació del saber en un mateix àmbit d'activitat, raó per la qual la terminologia representi el saber.

En relació amb el que comenta Cabré, hi ha estudis centrats en aspectes que s'allunyen de la teoria tradicional de la terminologia, estimulats pel

processament automàtic dels termes en els textos. Entre aquests estudis hi ha els de Daille *et al.* (1996), Jacquemin (2001), Tartier (2001) o Yoshikane *et al.* (1999), que analitzen les variants morfològiques o sintàctiques dels termes en relació amb el processament automatic dels termes i les seves variants. I Temmerman (2000), Pearson (1998) o Meyer i Mackintosh (2000) analitzen els termes en ús.

Així mateix, en aquest procés de revisió de la teoria i la pràctica terminològica es defineixen nous plantejament teòrics que ajuden a consolidar una nova manera d'afrontar el treball terminològic: la Socioterminologia, posicionament defensat per autors de França i de la part francòfona del Canadà, com Gaudin, Boulanger i Gambier; la Teoria Sociocognitiva de la Terminologia (TST) de Temmerman, la Teoria Comunicativa de la Terminologia (TCT) de Cabré i la Terminologia Textual de Bourigault i Slodzian. Seguidament descrivim cada un d'aquests models teòrics.

2.2.3 Socioterminologia

La socioterminologia és un plantejament que centra l'atenció en els aspectes socials i ideològics de les unitats terminològiques i té en compte de quina manera funcionen en un context social i són portadores de valors socials.

El terme *socioterminologia* és encunyat l'any 1981 per Boulanger, i també usat per Lerat i Slodzian, dins el marc de la planificació lingüística, ja que és un moment en el qual es produeix una sociologització de les diferents branques de la lingüística. Reprès i afinat per Gambier (1991), aquest terme comença a fer un veritable recorregut vers l'any 1986 en el col·loqui sobre la fertilització terminològica de les llengües romàniques en el qual participa Gambier (1987), i es consolida als anys noranta amb els treballs de recerca fets a la Universitat de Rouen, inspirats per Guespin i seguits per Gaudin, i també a la Universitat de Turku, de la mà de Gambier.

L'any 1994 el terme *socioterminologia* queda recollit en el *Dictionnaire de linguistique et des sciences du langage* (Justeson i Katz, 1995), publicat per Larousse.

> "La socioterminologie veut prendre en compte les aspects sociolinguistiques de la communication scientifique et technique. [...] Elle travaille le terme technique dans une optique qui part du signe linguistique. [...] La socioterminologie s'intéresse aux pratiques institutionelles qui visent l'observation, l'enregistrement et la normalisation des pratiques langagières dans les procès technologiques."

L'any 1998 Faulstich (1998) aporta una nova definició per a socioterminologia.

> "La socioterminologie est une discipline qui s'intéresse au mouvement du terme dans les langages de spécialités."

I l'any 2000 el terme és definit en el *Diccionario de organización y representación del conocimiento* de Barité (2013).

> "Socioterminología. 1. Rama de la Terminología que se ocupa del análisis de los términos (surgimiento, formación e interrelaciones), considerándolos desde una perspectiva lingüística en la interacción social. // 2. Disciplina eminentemente práctica del trabajo terminológico, que se fundamenta en el análisis de las condiciones sociales y lingüísticas de circulación de los términos."

I encara hi ha la definició que aporta Diki-Kidiri (2000) l'any 2000 a partir dels treballs que realitza sobre la planificació terminològica de les llengües africanes.

> "[La socioterminologie] s'est donné comme objectif d'étudier comment les locuteurs (utilisateurs, sujets, etc.) réagissent aux termes techniques, les utilisent ou les rejettent, et ce que cela induit comme relation de communication, et comme jeu et enjeu de pouvoir."

Aquest nou plantejament teòric sobre terminologia apareix sota una doble influència: la sociolingüística teòrica i la sociolingüística aplicada i es fixa com a objectiu "*l'étude de la circulation des termes en synchronie et en diachronie, ce qui inclut l'analyse et la modélisation des significations et des conceptualisations*" (Gaudin, 2003).

Així mateix, en aquest plantejament teòric l'estudi del vessant social de la terminologia representa una necessitat de renovar les estratègies amb què els investigadors han de resoldre nous reptes terminològics i repensar les pràctiques lingüístiques, gràcies a l'accés a la tècnica, la tecnologia i la ciència que hi ha hagut en els darrers anys.

La socioterminologia participa d'una reflexió central en tot el procés de la planificació lingüística. I és que el terme existeix realment en el discurs, i qualsevol discurs té lloc en el si d'un grup social. Gaudin proposa una visió àmplia de la terminologia: vinculant les pràctiques lingüístiques de caràcter especialitzat amb les pràctiques socials del mateix tipus mostra una ciència fusionadora. És una proposta que serveix per a construir un model lingüístic que estableix un lligam entre la part social (les condicions de la pràctica i la recepció) i la terminologia clàssica (el terme, la noció i la recerca associada) (Gaudin, 1993).

Concretament, la socioterminologia s'allunya de la prescripció wüsteriana i aposta per un enfocament descriptiu que tingui en compte la terminologia en l'ús real de la llengua, és a dir, en el medi natural on es troba i constantment canvia. L'enfocament renovador d'aquesta proposta se centra fonamentalment en quatre punts (Gómez González-Jover, 2007):

- Rebutja la idea inicial de la monosèmia i incorpora l'estudi de la sinonímia i la polisèmia.
- S'oposa a la compartició del coneixement en àmbits d'especialitat tancats i propugna un *contínuum* entre les ciències.
- S'allunya de l'estudi sincrònic de la llengua d'especialitat per a reintroduir el concepte d'història i de diacronia en la terminologia.

- Considera imprescindible incorporar l'estudi de la producció oral en la recerca terminològica i rebutja la idea que els textos escrits són els únics canals que garanteixen la comunicació científica i tècnica.

En aquest sentit, una orientació socioterminològica permet actualitzar els coneixements que fan referència al funcionament discursiu i social dels termes, enfocament que havia estat ignorat en la teoria tradicional. No es tracta de construir una crítica de la terminologia propugnada per Wüster, sinó d'ampliar el camp de la terminologia anant més enllà dels postulats idealistes i el seu voluntarisme i logicisme, en el si d'una perspectiva que té en compte coneixements diversos, des de la recuperació automàtica i les implicacions que té en el terreny de la descripció lingüística fins als avenços de la ciència i l'epistemologia (Gaudin, 2003).

2.2.4 Teoria Sociocognitiva de la Terminologia

La socioterminologia cognitiva també se centra en el vessant cognitiu de la terminologia des del punt de vista social, descartant la universalitat conceptual i introduint la noció de prototip per a explicar les diferents conceptualitzacions del mateix objecte segons patrons culturals i socials.

La Teoria Sociocognitiva de la Terminologia (TST), plantejada per Temmerman (2000), se centra en el vessant cognitiu de la terminologia des d'un punt de vista social. Aquesta teoria planteja certes divergències respecte als plantejaments de la Teoria General de la Terminologia exposada per Wüster. Concretament, de la teoria de Wüster no comparteix els aspectes següents (Faber Benítez, 2009):

- Els conceptes tenen un paper central respecte altres designacions lingüístiques.
- Els conceptes i les categories tenen uns límits clars.
- Les definicions terminogràfiques han de ser sempre intensionals.

- La referència monosèmica és una regla ("rule") en terminologia, en la qual hi ha una correspondència un a un entre termes i conceptes.
- El llenguatge especialitzat solament pot ser estudiat sincrònicament.

Temmerman propugna que aquestes premisses no són vàlides i exposa els principis següents:

- La llengua no pot considerar-se separadament dels conceptes pel fet que juguin un paper central en la concepció de les categories.
- Moltes categories tenen límits difusos i no poden ser clarament definides.
- Una definició òptima no pot ser limitada a un sol mode i, en última instància, depèn del concepte que hagi de ser definit.
- La polisèmia i la sinonímia apareixen freqüentment en el llenguatge especialitzat, i han de ser considerades en una anàlisi terminològica objectiva.
- Les categories, els conceptes i també els termes evolucionen amb el temps i han de ser estudiats diacrònicament. Per aquest motiu, els models cognitius juguen un paper important en el desenvolupament de noves idees.

Els principis expressats per Temmerman constitueixen el punt de partida d'una terminologia sociocognitiva, en la línia del que es planteja amb la Socioterminologia (Gaudin) i la Teoria Comunicativa de la Terminologia (Cabré).

L'aportació que fa la Teoria Sociocognitiva de la Terminologia és l'èmfasi en l'organització conceptual i, concretament, l'estructura de la categoria, enfront de les aproximacions de la lingüística cognitiva. Una altra aportació significativa d'aquesta teoria és la consideració que fa de la dimensió

històrica o diacrònica dels termes, en contraposició a l'anàlisi exclusivament sincrònica dels termes que recull la Teoria General de la Terminologia.

Recentment, el vessant sociocognitiu de la terminologia se centra en les ontologies, amb l'objectiu de poder implementar les representacions conceptuals. La combinació de terminologia i ontologia és anomenada *termontografia* (Temmerman i Kerremans, 2003), una aproximació multidisciplinar en la qual les teories i els mètodes per a una anàlisi terminològica multilingüe es combinen amb mètodes i directrius per a una anàlisi ontològica . D'aquesta manera, Temmerman s'acosta al plantejament fet prèviament per Meyer *et al.* (1992) als anys noranta, en el qual propugna que les bases de dades terminològiques són més rendibles si són organitzades d'una manera semblant a com estan organitzats els conceptes en la ment.

> "[...] term banks would be more useful, and useful to a wider variety of people, eventually even machines, if they contained a richer and more structured conceptual (i.e. knowledge) component than they do at present."

Segons el plantejament de Meyer, quan els termes formen part d'una base de dades terminològica de coneixement, hi ha una millora en les dades, pel fet que els conceptes i les designacions estan vinculats per mitjà de relacions significatives. Resulta un complement que enriqueix els resultats de l'estructura del coneixement ("causa-efecte", "objecte-funció") i també acaba de completar les relacions tradicionals ("genèric-específic", "part-tot").

En els darrers anys la termontografia s'ha anat consolidant fins al punt de tenir una entitat pròpia, més enllà de la terminologia sociocognitva, gràcies al desenvolupament de les tècniques d'enginyeria del coneixement i als processos de creació d'ontologies (Kerremans *et al.*, 2005).

2.2.5 Teoria Comunicativa de la Terminologia

La Teoria Comunicativa de la Terminologia (TCT) (Cabré, 1999b) presenta una profunda renovació dels postulats teòrics tradicionals plantejats per Wüster, els quals no permeten descriure la complexitat del lèxic especialitzat, ni explicar la comunicació especialitzada i les unitats més representatives, ni les varietats terminològiques.

L'observació de les dades terminològiques en el discurs natural, variat pel que fa a adequació de registres funcionals de la comunicació especialitzada, indica que són menys sistemàtiques, menys unívoques i menys universals que les dades observades per Wüster en el seu corpus normalitzat. Així, en el discurs especialitzat oral i escrit, la terminologia és un recurs expressiu i comunicatiu i, d'acord amb aquestes variables, el discurs presenta redundància, variació conceptual i variació sinonímica, i permet constatar que no sempre es produeix una perfecta equivalència entre llengües.

Segons Cabré, convé dissenyar un model teòric flexible i obert, que descrigui tota la complexitat conceptual de les unitats terminològiques, poder-les relacionar amb les altres unitats que les envolten i situar-les en una teoria interdisciplinària. Concretament, que sigui un model que tingui en compte la multidisciplinarietat de les unitats terminològiques (representativa, cognitiva i funcional) i la seva poliedricitat; la doble funció que tenen en el llenguatge especialitzat (representativa i comunicativa); la distinció entre el valor descriptiu i el valor prescriptiu que tenen; la variació conceptual inherent a qualsevol unitat de coneixement, vinculada a una cultura específica que determina la visió del món; la dependència lingüística de les unitats terminològiques i la variació denominativa inherent al discurs i a la comunicació, tant general com especialitzat, segons les característiques pragmàtiques del discurs.

L'aportació feta per Cabré en la TCT cerca nous fonaments que donin llum a una nova teoria sobre els termes basada en els fonaments del llen-

guatge i en el seu caràcter sociocultural i que donin compte dels termes com a unitats singulars i alhora semblants a altres unitats de comunicació dins un esquema global de representació de la realitat, admetent la variació conceptual i denominativa i tenint en compte la dimensió textual i discursiva dels termes.

Concretament, la TCT (Cabré, 1999b, 2002, 2003, 2010) es defineix pels paràmetres següents:

- La terminologia és una matèria que té un caràcter intrínsecament interdisciplinari, la qual rep les aportacions d'una teoria del llenguatge, que inclou aspectes pròpiament lingüístics, cognitius i socials; una teoria de la comunicació, i una teoria del coneixement.

- L'objecte d'estudi són les unitats terminològiques pròpiament dites, motiu pel qual és una teoria dels termes i no una teoria de la terminologia. Es consideren unitats terminològiques les que tenen caràcter lingüístic i que es produeixen en el si del llenguatge natural. I el seu caràcter específic radica en els aspectes pragmàtics i el mode de significació que tinguin. El marc en què s'han de descriure les unitats terminològiques concebudes com a unitats de naturalesa interdisciplinària (cognitives, lingüístiques i socials), ha de ser prou ampli i flexible. El caràcter de terme d'aquestes unitats s'activa segons l'ús que se'n faci en un context i una situació determinats.

- Els termes són unitats lèxiques que consten de forma o denominació i significat o contingut. La forma és constant, però el significat varia segons el tipus de situació i l'àmbit en què es troba.

- Els termes són unitats de forma i contingut en els quals el contingut és simultani a la forma. El contingut d'un terme mai no és absolut, sinó relatiu, segons cada àmbit i situació d'ús.

- Els conceptes d'un mateix àmbit especialitzat mantenen entre si relacions de diferents tipus. El conjunt d'aquestes relacions entre els conceptes constitueix l'estructura conceptual d'una matèria.

- El valor d'un terme s'estableix pel lloc que ocupa en l'estructura conceptual d'una matèria. Els termes no pertanyen a un àmbit, sinó que són usats en un àmbit amb un valor singularment específic.
- L'objectiu de la terminologia aplicada és recopilar unitats de valor terminològic en un tema i situació determinats i establir-ne les característiques.
- La finalitat aplicada de la recopilació i anàlisi de les unitats de valor terminològic usades en un àmbit és molt diversa i permet moltes aplicacions. I en totes s'activa la doble funció dels termes: la representació del coneixement especialitzat i la seva transferència. Els termes són usats en la comunicació especialitzada, caracteritzada per factors de tipus lingüístic i pragmàtic, la qual admet nivells d'especialització diferents, graus d'opacitat cognitiva variats, índexs diversos de densitat cognitiva i terminològica i propòsits diferents.

A partir dels supòsits anteriors, Cabré formula una proposta de construcció teòrica que anomena *model de les portes* (Cabré, 1999c, 2002), model en el qual insereix la terminologia en un marc d'entrada multiaccés que permet descriure les unitats terminològiques com a unitats lingüístiques i semiòtiques, cognitives i comunicatives. Com a unitats lingüístiques, perquè les unitats terminològiques són signes lingüístics, pertanyen a les llengües naturals, formen part de llurs gramàtiques i són descrites per mitjà de les mateixes propietats, estructures i condicions que les unitats lingüístiques. Com a unitats cognitives, perquè els termes vehiculen la representació de la categorització de la realitat que fan les especialitats. Com a unitats comunicatives, perquè els termes serveixen perquè els experts es puguin comunicar, i també per a formar nous experts i divulgar el coneixement especialitzat. Considerant l'entrada per la porta lingüística, Cabré fa l'aportació següent:

- La descripció de les unitats terminològiques es fa per mitjà dels textos o produccións lingüístiques orals i escrites dels especialistes en diferents situacions de comunicacions.

- Les unitats terminològiques són unitats prototípiques per a la representació eficient del coneixement especialitzat.

- Són unitats denominatives i designació que presenten variació (polisèmia i sinonímia).

- Les unitats terminològiques comparteixen amb altres unitats lingüístiques (morfològiques, sintagmàtiques i sintàctiques) l'expressió del coneixement especialitzat.

- Les unitats terminològiques es distingeixen per correspondre a unitats lèxiques, que ocupen un node pertinent en l'estructura conceptual d'una matèria i, semànticament, són les mínimes unitats autònomes en aquesta estructura.

- Les unitats terminològiques en una teoria del llenguatge natural no es conceben com a unitats separades de les paraules, sinó com a *valors* especialitzats de les unitats lèxiques contingudes en el lexicó del parlant.

- Una unitat lèxica no és en si ni terminològica ni no terminològica, sinó que per defecte és una unitat general que pot adquirir *valor especialitzat* o *terminològic* quan per les característiques pragmàtiques del discurs s'activa el seu significat especialitzat.

- Qualsevol unitat lèxica és en potència una unitat terminològica, encara que mai hagi activat aquest valor. Aquesta opció permet explicar els processos de *terminologització* i *desterminologització*.

- El sentit especialitzat descrit com a *valor* associat a les unitats del lèxic no és un conjunt predefinit i encapsulat d'informació, sinó una tria específica de característiques semàntiques segons les condicions de cada situació d'ús.

2.2.6 Terminologia textual

Bourigault i Slodzian (1999) exposen un nou enfocament teòric de la terminologia que es basa en una anàlisi completa dels nous objectius i pràctiques, teòriques i metodològiques, de la terminologia. Amb la demanda de terminologia per part de les empreses i les institucions neixen nous productes terminològics, que són complementaris a les bases de dades multilingües clàssiques i que s'adapten a les noves eines presents en les empreses.

Una mostra de nous productes terminològics que han aparegut en els darrers anys són els següents:

- Tesaurus per als sistemes d'indexació automàtica
- Terminologies de referència per als sistemes d'ajuda a la redacció
- Ontologies per als sistemes de presa de decisions
- Lèxics especialitzats per als motors de cerca temàtica a la xarxa
- Glossaris de referència i llistes de termes per a les eines de comunicació
- Índexs estructurats per a la documentació electrònica

Així mateix, Bourigault i Slodzian constaten el fet que hi ha variabilitat de les terminologies, en el sentit que no hi ha una única terminologia que representa el coneixement d'un àmbit d'especialitat. Les terminologies varien en nombre d'unitats i llur descripció segons l'eina que sigui utilitzada. Aquesta constatació, doncs, posa en dubte la universalitat de les terminologies. I és que una terminologia elaborada per una eina en un determinat moment no és mai idèntica a la que construeix una altra eina, motiu pel qual és fonamental el reaprofitament de terminologies.

Tenint en compte les constatacions anteriors, Bourigault i Slodzian propugnen un canvi en la pràctica terminològica, i és que la construcció d'una terminologia és una tasca d'anàlisi de corpus textuals. I paral·lelament

demanen una renovació teòrica de la terminologia, les bases de la qual se situen en la lingüística textual.

Bourigault i Slodzian aporten dues raons fonamentals al fet que la construcció d'una terminologia és una tasca d'anàlisi de corpus textuals:

1. Les applications de la terminologie sont le plus souvent des applications textuelles (traduction, indexation, aide à la rédaction) ; la terminologie doit "venir" des textes pour mieux y "retourner". C'est parce qu'elle n'est jamais déliée du texte qu'on parle de "terminologie textuelle".

2. C'est dans les textes produits ou utilisés par une communauté d'experts, que sont exprimées, et donc accessibles, une bonne partie des connaissances partagées de cette communauté, c'est donc par là qu'il faut commencer l'analyse.

En aquest sentit, propugnen que la tasca d'anàlisi terminològica ha de ser compartida entre l'expert d'un àmbit d'especialitat i un analista (lingüista terminòleg o cognitivista), perquè hi hagi un equilibri de punts de vista a l'hora de valorar la terminologia pròpia d'un àmbit d'especialitat. Per cada unitat seleccionada, l'analista, d'una banda, construeix una significació (*type*) a partir dels sentits (*occurrences*) recollits en el corpus. Aquesta tasca la fa a partir de la informació del corpus i de l'aplicació que ha de tenir. D'altra banda, l'expert és qui valida les descripcions construïdes per l'analista. Així mateix, consideren que la tasca d'anàlisi de corpus, per volum i terminis, s'ha de fer amb eines de la terminologia textual: eines de concordança, extractors de candidats a terme, extractors de relacions de candidats, classificadors, etc.

En aquest nou marc de treball terminològic, Bourigault i Slodzian fan una proposta teòrica i metodològica a partir de l'anàlisi de les noves pràctiques de la terminologia.

1. Proposition 1 : objet empirique d'une linguistique textuelle, le texte est le point de départ de la description lexicale à construire. On va du texte vers le terme. Les bases théoriques de la terminologie doivent être ancrées dans une linguistique textuelle.

2. Proposition 2 : le terme est un construit. Il est le produit d'un travail d'analyse, mené par le linguiste terminologue, dont les choix sont guidés par une double de contrainte de pertinence:

   – Pertinence vis-à-vis du corpus. Il s'agit de retenir et de décrire des structures lexicales qui présentent des caractéristiques à la fois spécifiques et stables. C'est à ce stade qu'intervient la validation par l'expert.
   – Pertinence vis-à-vis de l'application. Les unités finalement retenues doivent l'être en fonction de leur utilité dans l'application visée, qui s'exprime en termes d'économie et d'efficacité. La validation est à chercher du côté des utilisateurs de l'application.

Des d'aquest punt de vista, la tasca de la descripció lèxica serveix per a fixar, estabilitzar i homogeneïtzar un sentit, el resultat del qual és el terme. I fan referència a *normalització* quan la comunitat d'experts ratifiquen els significats com a termes d'un àmbit d'especialitat. L'enfocament de la terminologia textual demana un gir important en la metodologia proposada per Wüster. L'aproximació textual és descriptiva, perquè analitza les unitats lèxiques en el corpus, i no pas de caràcter normatiu, ja que l'aposta per la planificació lingüística està separada del treball terminològic. Amb l'estudi de les pràctiques textuals reals, la lingüística de corpus permet l'accés a expressions lingüístiques concretes de les quals serà possible extreure, i després normalitzar, els termes.

A tall de síntesi, per a tenir una visió general dels models teòrics que hem descrit en el present capítol, seguidament presentem un recull de les principals idees que propugna cada un d'aquests plantejaments teòrics.

RESUM DELS PLANTEJAMENTS TEÒRICS

Teoria tradicional de la terminologia (Wüster, 1930-1980)
Terminologia: matèria autònoma.
Objecte d'estudi: concepte.
Terme: unitat que designa un concepte.
Estudi del terme: normalització conceptual, denominativa.
Objectiu de la normalització: precisió i univocitat.
Estudia la sincronia del terme.

Socioterminologia (Boulanger, Gambier, Gaudin, 1981)
Anàlisi del terme: lingüística i interacció social.
Estudia la terminologia segons l'ús real de la llengua.
Rebutja la monosèmia.
Estudia la sinonímia i la polisèmia.
Proposa un *continuum* entre les ciències en lloc del coneixement en àmbits tancats.
Considera la diacronia en la terminologia.

Teoria Sociocognitiva de la Terminologia (Temmerman, 2000)
Estudia el vessant cognitiu de la terminologia des d'un punt de vista social.
Concep conjuntament la llengua i el concepte.
Considera la polisèmia i la sinonímia.
Estudia diacrònicament categories, conceptes i termes.

Teoria Comunicativa de la Terminologia (Cabré, 1999)
Terminologia: caràcter interdisciplinari (teories del llenguatge, de la comunicació i el coneixement).
Objecte d'estudi: unitats terminològiques (cognitives, lingüístiques i socials).
Terme: unitat lèxica que consta de forma i significat.
Concep el valor d'un terme pel lloc que ocupa en l'estructura conceptual.
Considera que els termes no pertanyen a un àmbit, sinó que són usats en un àmbit amb valor específic.

Terminologia Textual (Bourigault i Slodzian, 1999)
Basada en la lingüística textual: del text al terme.
Terme: pertinença al corpus i a l'objectiu de treball.

2.3 Recapitulació

En la primera part d'aquest capítol hem repassat les diferents accepcions que pren la denominació *terminologia* en el si del treball terminològic. De les accepcions descrites, en la nostra proposta experimental fem referència explícitament al conjunt de termes que representen els conceptes d'un àmbit d'especialitat, segons indiquen Rondeau (1984), Felber (1984) i Sager (1990). Així mateix, recollim la definició de *terminologia* que proposa el Comitè Tècnic 37 de la International Organization for Standardisation en la norma ISO 1087-1:2000 amb relació al treball terminològic. Concretament, és considerada com a ciència que estudia l'estructura, la formació, el desenvolupament, l'ús i la gestió de la terminologia en diverses àrees temàtiques i també com a conjunt de denominacions que pertanyen a una llengua d'especialitat. Aquesta segona accepció la prenem en consideració en el nostre treball en fer referència a terminologia.

A continuació identifiquem quines són les consideracions que convé tenir en compte respecte *concepte* i *terme*. Segons la norma UNE-ISO 1087-1:2009, *concepte* fa referència a una "unitat de coneixement formada per una combinació única de característiques", entenent *característica* en el sentit de "propietat abstracta d'un objecte o d'un conjunt d'objectes". En referència a *terme* hem recollit diverses consideracions, en les quals s'observa una posició de consens respecte al fet de considerar que un terme és una unitat que s'esdevé en un àmbit d'especialitat. Sobre el tipus d'unitat, hi ha una posició que considera que és una combinació entre denominació i noció (significant i significat o part conceptual i part lingüística) i n'hi ha una altra que considera que és una etiqueta (símbol lingüístic). Si ens centrem en el nostre estudi, la referència que fem de *terme* s'ajusta a la definició de la norma ISO 704:2009, que considera un terme com a "designation consisting of one or more words representing a general concept in a special language in a specific subject field".

Completem la primera part del capítol fent una descripció de quins són els límits entre *terme* i *paraula*. Hem constatat que hi ha posicions diverses al respecte. Així, Rondeau (1984) indica que són diferents sense assenyalar per què; Felber (1983) diu que les paraules poden tenir molts sentits, els quals són determinats pel context, i que els termes són símbols lingüístics assignats a un concepte o a més d'un, i Sager (1990) sosté des d'un punt de vista cognitiu que els ítems d'una disciplina que tenen una referència general són paraules i els que tenen una referència especial són termes, i des d'un punt de vista lingüístic manté que el límit es troba en l'ús que facin de paraules i termes especialistes i no especialistes d'una matèria. Posicions més recents com la de Slodzian (2000) manifesten que s'ha produït una "desespecialització" dels textos per l'accés generalitzat a la informació, motiu pel qual es desdibuixa la frontera entre paraula i termes; Cabré (2000) considera que termes i paraules formen part d'una mateixa unitat, que és la unitat lèxica, i que els termes no són unitats autònomes que constitueixin un lèxic especialitzat separat d'un lèxic general; Sager (2000) constata que els termes són unitats lèxiques que tenen una significació i una referència més precises que les de les paraules, i des d'un punt de vista comunicatiu Costa (2006) i Kocourek (2001) diuen que un terme es comporta com qualsevol altra unitat lèxica. Les diferents posicions descrites respecte a quina és la frontera entre terme i paraula han sofert una important evolució en els darrers anys amb l'admissió de la polisèmia del terme i la necessitat de disposar del context per a desambiguar-ne el sentit. Aquesta evolució va quedar ben palesa en el nostre projecte de recerca a l'hora de seleccionar quines unitats extretes del corpus corresponen a termes o a paraules de la llengua general i la diferent consideració que van mostrar els especialistes al respecte. Justament aquesta inflexió en el límit que hi ha entre terme i paraula ens ha fet considerar per a la proposta experimental que presentem en els capítols 4 i 6 l'ús de termes seleccionats per terminòlegs per tal d'avaluar els resultats obtinguts, i així evitar el que Sager (1998) comenta a l'hora d'establir la frontera entre paraules i termes, que "it can happen that non-specialists consider a word to be a term which is, however, only a general word for

the specialist; equally, it can happen that specialists use terms which their non-specialist audience take to be words in the general language" i que "the possibility of many lexical units to function both as words and as terms may even be a question of individual choice and interpretation of the speaker and listener".

En la segona part del capítol fem una descripció de quins són els principals plantejaments teòrics entorn de la terminologia. Així, la Teoria tradicional de la terminologia plantejada per Wüster va servir per a establir les bases teòriques de la terminologia. Els nous models teòrics que han sorgit a partir del període d'expansió de la terminologia (1975) han servit per a ampliar els límits establerts imposats per Wüster en el seu model teòric. L'evolució dels postulats que va establir Wüster s'ha produït amb el plantejament de la Socioterminologia, que situa la terminologia en un terreny social i ideològic, proposa l'estudi dels termes des d'un punt de vista diacrònic i sincrònic, i fa una aposta per un enfocament descriptiu que tingui en compte la terminologia en l'ús real de la llengua.

Un altre plantejament que trenca amb la teoria tradicional de Wüster és el de la Teoria Sociocognitiva de la Terminologia, que se centra en el vessant cognitiu de la terminologia des d'un punt de vista social. Fa èmfasi en el fet que la llengua no pot considerar-se separadament dels conceptes, que moltes categories tenen límits difosos, que la polisèmia i la sinonímia apareixen amb freqüència en el llenguatge d'especialitat i que tant les categories com els conceptes i els termes evolucionen amb el temps, motiu pel qual han de ser estudiats diacrònicament.

Amb la Teoria Comunicativa de la Terminologia es fa una profunda renovació dels postulats plantejats per Wüster. Es tracta d'un model teòric que té en compte la multidisciplinarietat de les unitats terminològiques (representativa, cognitiva i funcional) i la seva poliedricitat, és a dir, les funcions representativa i comunicativa en el llenguatge especialitzat; la variació conceptual, vinculada a una cultura específica que determina la visió del món; la dependència lingüística de les unitats terminològiques

i la variació denominativa inherent al discurs i la comunicació. Aquest plantejament teòric propugna un marc d'entrada multiaccés a la terminologia, que permet descriure les unitats terminològiques com a unitats lingüístiques i semiòtiques, cognitives i comunicatives.

La Terminologia textual exposa un plantejament teòric basat en una anàlisi dels nous objectius i pràctiques de la terminologia. Es propugna un canvi en la pràctica de la terminologia, basada en l'anàlisi de corpus textuals compartida entre un expert d'un àmbit d'especialitat i un analista (lingüista, terminòleg o cognitivista).

L'àmplia descripció dels diferents models teòrics que han estat plantejats en els darrers anys ens ha permès constatar que hi ha posicions ben allunyades respecte a la consideració de la terminologia com a disciplina. En el present treball assumim uns determinats supòsits teòrics amb relació a la terminologia i al seu objecte d'anàlisi, seguint l'aportació de Cabré (1999b) feta en la Teoria Comunicativa de la Terminologia. Concretament, prenem en consideració els aspectes següents:

- La terminologia com una matèria que té un caràcter intrínsecament interdisciplinari, la qual rep les aportacions d'una teoria del llenguatge, que inclou aspectes pròpiament lingüístics, cognitius i socials; una teoria de la comunicació, i una teoria del coneixement. En el nostre treball ens centrem en els aspectes lingüístics de la teoria del llenguatge.

- L'objecte d'estudi són les unitats terminològiques pròpiament dites. Es consideren unitats terminològiques les que tenen caràcter lingüístic i que es produeixen en el si del llenguatge natural. El caràcter de terme d'aquestes unitats s'activa segons l'ús que se'n faci en un context i una situació determinats, aspecte que considerem fonamental per a determinar el caràcter terminològic dels candidats a terme extrets d'un corpus especialitzat.

- Els termes són unitats lèxiques que consten de forma i significat. La forma és constant, però el significat varia segons el tipus de situació i l'àmbit en què es troba.
- Els termes són unitats de forma i contingut en els quals el contingut és simultani a la forma. El contingut d'un terme mai no és absolut, sinó relatiu, segons cada àmbit i situació d'ús.
- El valor d'un terme s'estableix pel lloc que ocupa en l'estructura conceptual d'una matèria. Els termes no pertanyen a un àmbit, sinó que són usats en un àmbit amb un valor singularment específic.
- L'objectiu de la terminologia aplicada és recopilar unitats de valor terminològic en un tema i situació determinats i establir-ne les característiques.
- La finalitat aplicada de la recopilació i anàlisi de les unitats de valor terminològic usades en un àmbit és molt diversa i permet moltes aplicacions. I en totes s'activa la doble funció dels termes: la representació del coneixement especialitzat i la seva transferència.

Capítol 3

EXTRACCIÓ AUTOMÀTICA DE TERMINOLOGIA

El processament del llenguatge natural (PLN) és una subàrea de la intel·ligència artificial que té com a objectiu el processament automàtic del llenguatge. En aquest sentit, el PLN és una disciplina que s'aplica a bastament en diversitat de tasques, entre les quals hi ha l'extracció d'informació, la recuperació d'informació, el resum automàtic de documents, la segmentació, la classificació de documents, la indexació de termes, la creació de plantilles, la manipulació de vocabularis controlats o diccionaris, la construcció d'ontologies, el reconeixement d'entitats amb nom (*name entity*) i el reconeixement automàtic de terminologia.

Centrant-nos en el reconeixement automàtic de la terminologia present en àmbits especialitzats, des de fa més de vint anys hi ha un interès creixent en aquesta àrea, en especial des del moment en què els sistemes de processament del llenguatge natural van passar "de l'etapa de desenvolupament a l'etapa d'aplicació" (Ananiadou, 1988, 1994b). En aquest sentit, l'extracció automàtica de terminologia es pot definir com el procés per mitjà del qual se selecciona d'un text o un conjunt de textos unitats candidates a ser termes fent servir mètodes computacionals (Oliver *et al.*, 2007; Foo, 2012).

Des d'un punt de vista metodològic, els estudis centrats en l'extracció automàtica de terminologia (EAT) se situen en l'àmplia categoria d'aproximacions de la lingüística computacional basades en corpus, categoria que s'ha anat consolidant en els darrers anys per una major accessibilitat als corpus i per la disponibilitat d'eines que permeten fer-ne una adequada explotació i consulta (Kageura i Umino, 1996). En aquest sentit, els corpus han esdevingut la font principal per al tractament automàtic de la llengua pel fet de poder tractar gran quantitat de dades textuals en suport electrònic (Condamines, 2005). Segons Kennedy (1998), durant més de tres dècades fent la tasca de compilació i anàlisi de corpus emmagatzemats en bases de dades ha nascut un nou corrent d'estudi anomenat *lingüística de corpus*. L'objectiu de l'anàlisi de corpus és de diversa mena: l'extracció de termes d'un corpus d'especialitat; l'adquisició o recuperació de coneixement morfològic, sintàctic o semàntic per a millorar el funcionament de les eines destinades al tractament automàtic de la llengua; l'extracció d'informació; la cerca d'informació; la millora dels sistemes de pregunta-resposta, o la millora de la traducció assistida per ordinador.

L'extracció automàtica de terminologia s'ha considerat cabdal en les darreres dècades pel fet que una terminologia ben construïda permet millorar tasques tan diverses com ara la traducció de textos; la construcció de diccionaris i vocabularis especialitzats; la redacció tècnica; la producció documental; la classificació, indexació i arxiu de documents; la recuperació d'informació (monolingüe i bilingüe); el reconeixement d'entitats amb nom; la creació de tesaurus; l'extracció, reorganització i reformulació de coneixement representat en un text; l'anàlisi documental, i també la mineria de textos o el resum de documents (Heid i McNaught, 1991; Frantzi i Ananiadou, 1997; Vu *et al.*, 2008).

Concretament, per a la tasca de *traducció de textos*, els termes són extrets dels documents d'origen i destinació i passen a formar part d'un diccionari o una base de dades terminològica que és aprofitada i ampliada en cada nova traducció.

En la tasca de *construcció de diccionaris especialitzats*, incorporar l'extracció automàtica de terminologia permet estalviar l'esforç humà que representa la compilació manual dels termes presents en els diccionaris. I és que disposar d'una àmplia cobertura de diccionaris especialitzats és cabdal per al processament del llenguatge natural, els quals són difícils de compilar manualment, especialment en àmbits tècnics, perquè constantment es creen nous termes que representen nous conceptes, fet que dificulta l'actualització constant de la terminologia (Utiyama *et al.*, 2000; Witschel, 2005).

L'àmbit de la *classificació de documents* té com a objectiu agrupar automàticament documents en unes categories predefinides. La majoria de tècniques de classificació seleccionen paraules per a representar característiques dels documents i classificar-los per similitud. Les tècniques d'extracció de termes multiparaula poden ajudar en la tasca d'obtenir més qualitat dels termes representatius dels documents (Montanes *et al.*, 2005; Hovy *et al.*, 2000).

En la tasca d'*indexació de documents textuals*, és possible assignar automàticament termes d'indexació a un text per a facilitar-ne la recuperació posterior. Les unitats lèxiques extretes representen els "conceptes" propis d'un document i són proposades com a candidats descriptors per als documents que s'han d'indexar.

Per a la tasca de *recuperació d'informació* és fonamental disposar dels termes rellevants d'una col·lecció de documents per a poder-ne indexar els continguts amb l'objectiu de guiar l'usuari en la seva cerca d'informació (Dias *et al.*, 2000) i també per a millorar la cerca i recuperació de continguts de la xarxa (Witschel, 2005); per aquest motiu, la recerca en aquest camp se centra en la indexació automàtica i l'extracció automàtica de paraules clau.

En la tasca de *creació de tesaurus*, les unitats lèxiques representen els conceptes d'un àmbit al qual pertanyen els textos (Amar i David, 2001). Els tesaurus s'empren per a donar suport als processos de cerca d'informació o per a la consulta de bases de dades. La constitució automàtica de tesaurus té una llarga tradició i els estudis fets en aquest àmbit fan referència a la descoberta de nous termes o a l'establiment de relacions semàntiques (Jacquemin, 1997).

La tasca de *reconeixement d'entitats amb nom* permet identificar les entitats de noms de persones, llocs i organitzacions de manera automàtica a partir de corpus i endreçar-los en la categoria corresponent. Les tècniques d'extracció de termes multiparaula poden ajudar en la localització d'entitats amb nom i el seu corresponent processament (Pal *et al.*, 2010).

El *resum automàtic* permet generar automàticament breus resums de documents. Aquesta tasca és molt important amb el creixent nombre de documents que hi ha a la xarxa i la necessitat de recuperar-ne el contingut. Tradicionalment és una tasca que s'ha dut a terme en documents ben estructurats, per ser més coherents i contenir frases i paràgrafs clau per a descriure les idees principals d'un text. Actualment també s'aplica a textos curts, no formals i no gaire ben estructurats (Dunning, 1993; da Silva *et al.*, 1999).

En l'àmbit de la documentació, la terminologia és percebuda com una *representació del contingut dels documents i la clau per a poder-hi tenir accés*. I per a la creació de recursos terminològics s'ha de disposar d'un gran nombre de documentació per a cobrir tant la representació del contingut com l'accés a aquest. És el punt en què s'estableix el lligam entre terminologia i documentació. La terminologia treballa amb gran diversitat de contextos, de punts de vista i varietat de temes, que donen lloc a un ampli ventall d'aproximacions i aplicacions que se situen en el camp de les ciències de la informació i de la comunicació (El Hadi, 2006). En aquest sentit, la terminologia intervé en l'elaboració de vocabularis de referència per a aplicacions de caràcter divers que permeten

l'accés a la informació: tesaurus per als sistemes d'indexació automàtica, índexs estructurats per a la documentació tècnica hipertextual, ontologies per a memòries d'empresa, terminologia per als sistemes d'ajuda en la presa de decisions o per als sistemes d'extracció d'informació. Així, doncs, els termes constitueixen l'objecte d'estudi per a la terminologia i per a la documentació per mitjà dels llenguatges documentals (Mustafa el Hadi, 2005).

Un factor afegit a la diversitat d'aplicacions que ofereix l'extracció automàtica de terminologia i que n'accentua la rendibilitat és l'aparició de les noves tecnologies. El tractament automàtic de la informació textual publicada a la xarxa ha esdevingut un aspecte clau. En aquest context informacional canviant, la terminologia hi juga un doble paper: d'una banda, posar límits als diferents àmbits de coneixement per a facilitar-ne la gestió per part dels usuaris i, d'altra banda, millorar el funcionament de les eines que permeten el tractament automàtic de la informació. La terminologia, doncs, des d'un punt de vista informacional és un recurs imprescindible per a la gestió de la informació, ja que permet descriure camps de coneixement igual que les nomenclatures, els tesaurus i les ontologies i intervé en diferents processos de gestió de la informació, especialment en el sector empresarial, com la gestió de coneixement, la intel·ligència econòmica, la posada en valor de la memòria econòmica d'una empresa o la gestió del patrimoni informacional de l'empresa. Des d'un punt de vista tecnològic, la terminologia és considerada tant un recurs lingüístic necessari per al bon funcionament de les eines de tractament automàtic de la llengües (traducció automàtica, traducció assistida per ordinador, alineació bilingüe i multilingüe d'expressions multiparaula, anàlisi de contingut, resum automàtic, cerca i filtratge d'informació) com el resultat produït per les eines de tractament automàtic de les llengües, terminologia que es construeix a partir de grans volums de dades textuals amb la finalitat de disposar de llargues llistes de termes per a elaborar diccionaris, tesaurus, entre altres (Chaudiron, 2005; Piao *et al.*, 2005). I també és una peça més dels motors de cerca, en els quals té una llarga aplicació, i és que la gran quantitat de coneixement lingüístic que s'aplica als sistemes de

cerca, sovint no percebuda per l'usuari mitjà, serveix perquè els usuaris finals puguin localitzar la informació amb la màxima exactitud per mitjà dels cercadors (Peñas *et al.*, 2001).

En les darreres dècades s'han emprat diversos mètodes destinats a automatitzar la identificació de termes presents en els diferents àmbits d'especialitat. Seguidament, fem una descripció dels principals mètodes estadístics, mètodes lingüístics i mètodes híbrids que són aplicats en extracció automàtica de terminologia.

3.1 Mètodes aplicats a l'extracció automàtica de terminologia

L'extracció automàtica de terminologia empra mètodes computacionals per a seleccionar d'un text o un conjunt de textos candidats a terme que puguin ser processats posteriorment per a dur a terme un treball terminològic [(Oliver *et al.*, 2007, p. 78); (Foo, 2012, p. 2)]. L'extracció manual de terminologia es distingeix de l'extracció automàtica de terminologia pel fet que l'extracció de termes és feta per persones. I quan es fa referència a *extracció de termes* es fa èmfasi únicament en l'extracció de termes en general, sense referir-se a la manera com es duu a terme la tasca (Foo, 2012, p. 17).

En Pazienza *et al.* (2005) es descriu quin hauria de ser el procés ideal de reconeixement terminològic basant-se en l'ús d'estratègies lingüístiques. Aquest procés passa per l'anàlisi de l'àmbit d'especialitat del corpus, la identificació com a mínim de la categoria gramatical de les unitats lèxiques, la identificació i extracció dels candidats a terme a partir de regles, la unificació de les variants en el terme original i la implementació de filtres lingüístics per a millorar els resultats obtinguts. Al final del procés, s'ha d'obtenir una llista de bons candidats a terme susceptibles de formar part de la llista final de termes. A partir d'aquí, un expert duu a terme la validació manual de les unitats lingüístiques seleccionades. Ara bé, la di-

ficultat de trobar informació aprofundida sobre el terme, que defineixi les propietats que caracteritza unívocament els termes, i la "traducció" d'aquestes característiques a un sistema d'una manera clara, juga un paper central en la recerca sobre lingüística computacional.

Encara que els termes continguts en documents especialitzats són molt precisos perquè tenen relació amb els conceptes que difonen els experts (Smadja, 1993; Bourigault i Jacquemin, 1999), no hi ha regles formals que defineixin les propietats que determinen automàticament què és un terme i què no ho és i tampoc la millor metodologia per a extreure'ls. A més, la variació terminològica incrementa la dificultat a l'hora de detectar automàticament un terme. D'entrada, els termes són monoreferencials (hi ha una correspondència d'un a un entre termes i conceptes), però a la pràctica s'han de tenir en compte les ambigüitats (un mateix terme té correspondència amb diversos conceptes) i les variants (molts termes corresponen al mateix concepte). Si hom té com a objectiu fer una recuperació i estructuració sistemàtica dels continguts d'un determinat àmbit de coneixement, llavors la variació terminològica s'ha d'afrontar com una part essencial de la mineria terminològica (Daille, 2003; Nenadic *et al.*, 2004). S'han desenvolupat pocs mètodes sobre variació terminològica: el sistema Fastr (Jacquemin *et al.*, 1997; Jacquemin, 1997) tracta les variacions morfològiques i sintàctiques a partir de meta-regles lexicalitzades que es fan servir per a la normalització terminològica, en què la variació semàntica queda resolta per mitjà de WordNet (Miller, 1995). De manera semblant, el mètode C/NC-Value emprat en extracció de terminologia (Frantzi *et al.*, 2000) tracta la variació ortogràfica, morfològica i estructural dels termes i els sinònims (Nenadic *et al.*, 2005).

En la dècada passada i en els darrers anys la recerca sobre terminologia computacional s'ha centrat en diverses estratègies per a extreure i reconèixer termes fent servir tècniques supervisades i no supervisades, amb l'objectiu de localitzar els termes més significatius en un corpus especialitzat. Algunes de les aproximacions al problema que s'han dut a terme (Pazienza *et al.*, 2005, p. 2) són les següents:

a) les mesures estadístiques proposen de definir el nivell de *termhood* -–nivell de pertinença a un àmbit d'especialitat— de la llista de candidats a terme, per exemple per a localitzar les mesures més apropiades que ajudin a triar bons termes d'una llista de candidats;

b) els terminòlegs computacionals han intentat definir, identificar i reconèixer termes, tenint en compte les propietats lingüístiques bàsiques, fent servir tècniques de filtratge lingüístic amb l'objectiu d'identificar patrons sintàctics i terminològics específics;

c) les aproximacions híbrides miren d'usar aquestes dues visions alhora, tenint en compte les metodologies lingüística i estadística per a reconèixer termes.

Paral·lelament a l'aplicació de diverses aproximacions que ajuden a identificar les unitats terminològiques, s'han incorporat les mesures de cobertura i precisió per tal d'observar el nivell de detecció de termes que aconsegueixen les eines que fan extracció de termes monoparaula i multiparaula, com varien els resultats per a termes resultants de diferents processos de formació terminològica, quina ràtio hi ha per als *types* de les formes dels termes en oposició als *tokens* o com de bé el sistema reconeix noves formes terminològiques que no són en el lexicó o prèviament trobades en corpus d'entrenament (Lauriston, 1995). La cobertura és definida com la ràtio dels elements correctament identificats dividida pel nombre total d'elements vàlids o adequats; mesura l'efectivitat del sistema. I la precisió d'un sistema de recuperació és definida com la ràtio dels elements correctament identificats dividida pel nombre total d'elements recuperats; mesura la qualitat del material recuperat (Salton, 1989).

Abans de continuar, però, convé definir què entenem per *terme multiparaula*, *token* i *type*.

terme multiparaula: a multi-word term (MWT) is an expression consisting of more than one word with a grammatical structure and a specific meaning (Huo, 2012, p. 1).

> **token**: a token is a contiguous sequence of characters (McInnes, 2004, p. 5).
>
> **type**: a type is the class of all tokens containing the same character sequence (Manning *et al.*, 2008, p. 22).

El resultat que s'obté de l'extracció de terminologia pot ser formalitzat com un problema de classificació –el terme *t1* és rellevant o irrellevant (Vivaldi *et al.*, 2001)– o bé com un problema de rànquing –el terme *t1* és més rellevant que no pas el terme *t2* (Cohen *et al.*, 1999)– i és el pas previ per a la construcció d'ontologies.

Tal com ja hem apuntat, els mètodes d'extracció d'automàtica de terminologia empren diferents estratègies per a assolir l'objectiu d'obtenir les unitats lèxiques que són representatives d'un corpus d'especialitat. Seguidament descriurem breument quins són aquests mètodes i els classificarem segons facin servir una estratègia estadística, lingüística o híbrida.

3.1.1 Mètodes estadístics

Els mètodes estadístics s'han convertit en referents gràcies al desenvolupament de la lingüística de corpus. Aquests mètodes utilitzen càlculs propis d'altres àrees com la recuperació d'informació i la detecció de col·locacions. Els mètodes estadístics tenen una funció destacada en el processament del llenguatge natural i ofereixen bons resultats en camps tan diversos com l'etiquetatge morfosintàctic (*part-of-speech tagging*) i l'alineació de segments (Daille, 1994).

En l'àmbit del processament del llenguatge natural, en qualsevol de les seves tasques, la freqüència és un càlcul estadístic destacat per a identificar relacions textuals rellevants i per a identificar unitats multiparaula, tal com recullen Smadja (1993), Daille (1995, 1997), McEnery *et al.* (1997), Merkel i Andersson (2000), Piao i McEnery (2001) i Pereira *et al.* (2004).

La freqüència d'aparició d'un terme en un corpus és una mesura estadística significativa en la tasca d'extracció automàtica de termes, ja que com més vegades aparegui el candidat en un context més indicis tindrem que sigui un terme. Així, els sistemes d'extracció de terminologia més simples processen els documents usant únicament la freqüència per a endreçar la llista de resultats obtinguts (Manning i Schütze, 2003). Ara bé, hi ha unitats lingüístiques que apareixen amb freqüència en un corpus i no són termes. En aquest sentit, la freqüència d'un terme no és un indicador perfecte per a la recuperació automàtica de termes.

Així mateix, la freqüència d'aparició d'un terme en un corpus és molt fàcil de comptabilitzar per part dels sistemes d'extracció automàtica de temes; amb tot, la variació terminològica introdueix dificultats a l'hora de fer aquest recompte, pel fet que en un text un terme hi pot aparèixer en la seva forma completa, en referències anafòriques (pronoms) o en noms que s'hi refereixen. Per aquest motiu, és difícil comptar el nombre exacte de vegades que apareix un terme en un corpus.

Els mètodes estadístics reconeixen les unitats terminològiques a partir de la freqüència que tenen en un corpus marcat temàticament. Malgrat ser un càlcul molt senzill, un altre problema que presenta és que no permet de recuperar amb facilitat termes que apareixen poques vegades en un corpus d'especialitat, mancança que es pot resoldre fent servir filtres lingüístics o bé mesures estadístiques que permeten millorar els resultats de la freqüència, com ara la mesura Log-likelihood (Daille, 1995). A més, tenir en compte solament la freqüència per a la recuperació d'unitats multiparaules pot introduir errors en el procés d'identificació d'unitats pròpies d'un determinat àmbit d'especialitat, ja que no sempre les combinacions de paraules que apareixen sovint en un corpus són susceptibles de ser considerades pròpies d'un àmbit d'especialitat (Frantzi i Ananiadou, 1996a), en aquest sentit hi ha termes complexos que són fruit de relacions lèxiques específiques que contenen mots poc freqüents, i també convé tenir en compte que com més paraules freqüents té un n-gram menys rellevant és (Dias *et al.*, 1999). Així, mentre que n-grams que formen part d'n-grams

més llargs introdueixen un factor negatiu pel que fa a la seva rellevància, les seqüències de paraules augmenten la seva probabilitat d'importància com més alt és el nombre d'n-grams llargs en el text, tal com recull Dias i Nunes (2004). Per *n-gram* considerem una seqüència contínua de paraules d'un text.

> "Ngrams are defined as a contiguous or non-contiguous sequence of words, often called *tokens*, that occur in some proximity to each other in a corpus. [...] The unique tokens are the unigrams (1-grams) of the corpus. A bigram (2-gram) is a sequence of two tokens in a corpus and a trigram (3-gram) as a sequence of three tokens" (McInnes, 2004, p. 5).

Una altra estratègia que fan servir els mètodes estadístics és mesurar el grau d'associació que hi ha entre alguns dels components d'un candidat a terme. Per a calcular el grau d'associació que hi ha entre els components d'un candidat a terme es basen en càlculs estadístics, que oscil·len des de simples freqüències fins a mesures més complexes. L'inconvenient que tenen aquests càlculs és que extreuen totes les associacions lèxiques possibles sense distingir si són termes multiparaula, col·locacions o combinacions casuals de paraules de la llengua. Per a reduir aquest inconvenient, s'incorpora informació lingüística *a priori*, tal com es fa en l'eina Xtract (Smadja, 1993) o l'eina Acabit (Daille, 2003).

Així mateix, les tècniques estadístiques són emprades per a la normalització de termes, com en el treball de Frantzi i Ananiadou, en el qual es proposa l'ús de la fórmula estadística C-value per a identificar termes relacionats com ara "soft contact lenses", "hard contact lenses", "contact lenses", alhora que és capaç de predir que la combinació "soft contact" no és pas un terme (Frantzi i Ananiadou, 1996b).

Per tal d'obtenir millors resultats en l'extracció de termes els mètodes estadístics permeten l'ús de llistes de paraules funcionals o buides (*stopwords*) –articles, pronoms, preposicions, conjuncions, etc.– per a evitar que al començament o al final del candidat a terme hi hagi una paraula buida de

contingut, i també l'ús de mesures d'associació lèxica per a poder extreure únicament els candidats que tenen més probabilitat de ser termes per grau d'associació, com ara la ràtio log-likelihood, la prova khi quadrat de Pearson, la ràtio Odds, el coeficient PHI, la prova *t* de Student, el coeficient Dice, la mesura informació mútual, entre altres (vegeu capítol 5).

Els mètodes únicament estadístics (Dunning, 1993; Dias, 2002) extreuen unitats multiparaula d'un corpus a partir de mesures estadístiques. Com que admeten corpus en format de text pla i solament necessiten la informació que hi ha en el text, aquests sistemes són molt flexibles i permeten extreure unitats multiparaula independentment de l'àmbit i la llengua del corpus. I encara que de manera ocasional facin servir filtres lingüístics, no necessiten lexicons ni regles.

Aquests mètodes depenen sobretot de la correcta identificació de bigrams per a començar el procés iteratiu, i solament poden identificar associacions textuals que hi ha en el context en què apareixen. Com a conseqüència d'això, unitats rellevants no poden ser introduïdes directament en una base de dades lèxica, ja que no garanteixen una determinada estructura lingüística per a aquest propòsit (Dias, 2003). A més, la pròpia natura dels mètodes estadístics fa que no siguin prou precisos a l'hora d'identificar unitats multiparaula en freqüències baixes, sobretot quan aquestes apareixen una o dues vegades en el corpus. Per aquest motiu, la recerca que es fa amb aproximacions estadístiques no té en compte la presència d'unitats multiparaula que apareixen amb poca freqüència en el corpus, quan en aquesta franja de resultats hi ha una proporció significativa d'unitats multiparaula, En conseqüència, la utilitat de les aproximacions únicament estadístiques en les aplicacions de processament del llenguatge natural és limitada (Piao *et al.*, 2005).

Si es treballa amb un corpus petit, aquest tipus de mètodes generen molt de silenci o un nombre de termes no reconeguts del total de termes presents en un text. Si el corpus és gran, sempre hi ha un nombre de termes que, per la seva baixa freqüència, no es poden recuperar. A més, també

generen soroll, és a dir, recuperen candidats a terme que no tenen valor terminològic, això es deu al fet que en els textos especialitzats també hi apareixen paraules amb significat no especialitzat que pertanyen a la llengua general i que hi apareixen amb freqüència elevada.

Una altra limitació que tenen els mètodes estadístics és que no permeten arribar a fer generalitzacions que contribueixin a explicar fenòmens del llenguatge general, ja que fan servir estratègies independents de la llengua; en canvi, els mètodes lingüístics sí que permeten fer aquest tipus de generalitzacions. Les limitacions que presenten els mètodes estadístics fa que s'hagin de combinar amb altres tipus d'estratègies per a millorar la detecció d'unitats especialitzades presents en un domini especialitzat.

RESUM DELS MÈTODES ESTADÍSTICS

- Usen la freqüència per a identificar els termes d'un corpus.
- Incorporen mesures d'associació lèxica.
- Mesuren el grau d'associació entre els components dels candidats a terme.
- Permeten flexibilitat en l'extracció de candidats, independentment de l'àmbit d'especialitat i la llengua.
- No usen lexicons ni regles lingüístiques.
- Tenen dificultat per identificar termes poc freqüents.
- En corpus petits, generen molt de silenci.
- En corpus grans, generen soroll.
- No fan generalitzacions lingüístiques, fan servir estratègies independents de llengua.

3.1.2 Mètodes lingüístics

Els mètodes lingüístics són usats ja en els primers treballs sobre extracció de terminologia. En treballs com els de Ananiadou (1988, 1994b,a) el focus d'estudi se centra en la morfologia lèxica, i també hi ha interès a desenvolupar metodologies per al reconeixement de termes que puguin aplicar plantejaments teòrics referents a la formació dels termes. I és que la lingüística teòrica se centra en l'estructura de la llengua general. Ananiadou dissenya un model integrat d'estructura de paraula i terme basat en els resultats d'una anàlisi de termes de l'àrea d'immunologia (àmbit de medicina) en anglès i en models que es troben en la bibliografia de llengua general (Selkirk, 1982; Mohanan, 1986). Concretament implementa un lexicó i una gramàtica morfològica computacional capaç de detectar compostos i altres formes simples i complexes d'una manera teòrica satisfactòria i, a més, demostra que millora el reconeixement de termes. La identificació en aquest nou nivell és una contribució a la teoria morfològica.

Així mateix, treballs com els de Daille (1997) i també Jacquemin (1999) fan una anàlisi sintàctica aprofundida dels candidats a terme, que inclou anàlisi morfològica i anàlisi de dependències (*head-modier dependency analysis*). Aquesta anàlisi tan acurada l'apliquen a diferents subàmbits especialitzats, en els quals s'observa una forta variació sintàctica (distribució d'etiquetes morfosintàctiques, patrons sintàctics, etc.), fet que afegeix dificultat a l'hora d'aplicar les especificacions gramaticals a diferents àmbits d'especialitat.

Les aproximacions lingüístiques de reconeixement terminològic miren d'identificar els termes tenint en compte les propietats sintàctiques que tenen (Pazienza *et al.*, 2005), l'ús de dades sintàctiques (Bourigault, 1992), l'ús de patrons lingüístics per mitjà dels quals els sistemes poden localitzar formes correctes, el coneixement lingüístic (Basili *et al.*, 1997) o l'ús d'expressions regulars per a identificar formes de candidats a terme (Daille, 1994; Justeson i Katz, 1995).

El coneixement lingüístic que fan servir aquests mètodes per a reconèixer termes es basa en recursos lexicogràfics, com diccionaris de termes o diccionaris de paraules auxiliars –Fastr (Jacquemin, 1994)–; recursos morfològics, com poden ser patrons d'estructura interna de la paraula –Terms (Justeson i Katz, 1995)–; recursos morfosintàctics, com patrons morfosintàctics –Termino (David i Plante, 1990)–; extracció de la màxima llargada de frases nominals basada en l'anàlisi gramatical superficial i anàlisi de la màxima llargada de frases nominals per a extreure unitats adequades de candidats a terme (elements que marquen la frontera exterior de la unitat terminològica) –Lexter (Bourigault *et al.*, 1996)–; ús de la categoria gramatical i alineació de paraules –Termight (Dagan i Church, 1994)–, o funcions sintàctiques –Nodalida (Arppe, 1995). I, esporàdicament, recursos semàntics, com ara classificació semàntica, i recursos pragmàtics, com poden ser representacions tipogràfiques o informació de disposició del terme en el text –Drouin (Drouin, 1997).

Les estratègies lexicosintàctiques dels sistemes que fan servir mètodes lingüístics són de dos tipus: d'una banda, les que fan servir solament la paraula, el lema i la categoria gramatical com Bourigault *et al.* (1996); Justeson i Katz (1995); d'altra banda, les que fan servir informació que prové de l'anàlisi sintàctica superficial per a la identificació dels termes com Arppe (1995); Hulth (2003). Ambdós tipus d'estratègies es basen en l'assumpció que els termes que tenen una freqüència alta en un corpus especialitzat corresponen a frases nominals i tenen com a tasca principal identificar aquestes frases nominals en els textos.

Els actuals sistemes de reconeixement automàtic de terminologia incorporen l'etiquetatge morfosintàctic (*part-of-speech tagging*) de les paraules del document, fet que permet categoritzar cada paraula en un *type* lèxic; apliquen expressions regulars basades en patrons de concordança, i apliquen filtres lingüístics per tal d'afinar els resultats obtinguts amb l'objectiu d'identificar els termes que hi ha en un document. Els filtres lingüístics s'incorporen per a seleccionar termes multiparaula a partir dels reconeixe-

ment d'uns certs patrons. L'ús de patrons lingüístics per a la identificació de termes té la seva limitació, perquè hi ha un nombre finit de patrons lèxics, els quals s'han fet servir de manera general i experimentat per a identificar termes (Arppe, 1995; Bourigault *et al.*, 1996; Gaizauskas *et al.*, 2000).

Concretament, l'eina Yatea (Aubin i Hamon, 2006), a partir d'un corpus segmentat per paraules i frases, lematitzat i etiquetat a nivell morfosintàctic, extreu frases nominals que s'assemblin a termes. Permet localitzar termes de grans corpus; no depèn d'una llengua específica, en el sentit que els patrons lingüístics poden ser adaptats o crear-ne de nous. L'estratègia principal de l'anàlisi dels candidats a terme es basa en l'explotació de l'anàlisi simple dels patrons i en desambiguació endògena, és a dir, l'explotació dels resultats intermedis d'extracció. La desambiguació exògena és possible per mitjà d'una llista de termes de referència. L'ús de terminologies influeix positivament en la identificació de les frases nominals, l'anàlisi i, finalment, l'extracció de la llista de termes.

I darrerament s'ha presentat TTC TermSuite (Rocheteau i Daille, 2011), el primer paquet d'eines destinat a l'extracció multilingüe de terminologia a partir de corpus comparables en anglès, francès, alemany, espanyol, letó, xinès i rus. L'arquitectura funcional de TTC TermSuite se centra en quatre passos: el preprocessament del text, l'anàlisi lingüística, l'extracció terminològica i l'alineació de termes. TTC TermSuite disposa d'una interfície gràfica d'usuari que millora l'anàlisi textual i lingüística destinada a l'extracció terminològica monolingüe i l'alineació terminològica bilingüe.

Els mètodes lingüístics tenen com a limitació la identificació de termes monolingües, ja que els sistemes necessiten tècniques molt especialitzades des d'un punt de vista lingüístic per a poder aïllar els possibles candidats a terme, com passa amb el sistema de Bourigault (Dias, 2003).

Així mateix, i de manera general, són mètodes que també generen molt de soroll, és a dir, proposen molts candidats a terme que després s'han de revisar manualment, i generen silenci, ja que no detecten totes les unitats candidates a terme, ja sigui perquè aquestes corresponen a patrons morfològics que no han estat recollits, per problemes en el procés de desambiguació o per deficiències del sistema mateix. A més, pel tipus de coneixement que fan servir, aquests mètodes solament són aplicables a una llengua. Per a traslladar-los a una altra llengua, cal fer un estudi lingüístic previ.

RESUM DELS MÈTODES LINGÜÍSTICS

- Identifiquen termes a partir de propietats sintàctiques, patrons lingüístics i expressions regulars.
- Fan servir coneixement lingüístic basat en recursos lexicogràfics, morfològics, sintàctics, anàlisi gramatical i superficial.
- Combinen etiquetatge morfològic, expressions regulars basades en patrons de concordança i filtres lingüístics.
- S'han d'adaptar a la llengua de la qual es fa l'extracció de candidats.
- Generen soroll i silenci per manca de patrons morfològics, errors de desambiguació o deficiència del sistema.

3.1.3 Mètodes híbrids

Els sistemes d'extracció de terminologia més recents combinen tècniques lingüístiques i estadístiques en les aproximacions híbrides. L'ús de mètodes que combinen aquestes dues tècniques permeten confirmar o rebutjar la condició de terme d'una unitat lingüística. Les tècniques estadístiques ofereixen informació relacionada amb l'ús dels mots, fet que supleix la competència pragmàtica que té un especialista envers un terme.

En els mètodes híbrids l'ordre d'aplicació del tipus de coneixement és important, ja que els resultats que s'obtenen són diferents. I és que la rendibilitat de les mesures estadístiques augmenta quan el coneixement estadístic s'aplica en la llista de candidats seleccionats prèviament a partir de filtres lingüístics, perquè la part lingüística ajuda a seleccionar els candidats admissibles abans d'aplicar els tests numèrics –Acabit (Daille, 1995), Clarit (Evans i Zhai, 1996). A partir d'aquí, les mesures seleccionen i endrecen els candidats d'acord amb la definició de *termhood* —nivell de pertinença a un àmbit d'especialitat-– o *unithood* —cohesió lèxica-– implementades en cada mesura específica. En canvi, els mètodes que apliquen primer coneixement estadístic i després coneixement lingüístic tenen problemes de silenci, com també passa amb els mètodes lingüístics (Drouin, 1997).

Un dels primers sistemes a usar aproximació híbrida fou Earl (1970). En el seu treball extreu primer les frases nominals com a candidats a terme i després les selecciona d'acord amb la freqüència en què apareixen en el corpus. Uns quants anys més tard, en el treball de Daille (1994) els candidats lingüístics obtinguts a partir de patrons sintàctics es filtren amb diferents mesures estadístiques com ara log-likelihood, informació mútua i freqüència. I en el treball de Justeson i Katz (1995) s'hi aplica un sistema semblant en el qual s'usen expressions regulars per a extreure els candidats lingüístics del corpus, els quals són endreçats després per freqüència.

Una estructura més complexa es planteja a Ana (Automatic natural acquisition of a terminology) (Enguehard i Pantera, 1995), que usa ambdós mètodes. Primer, extreu termes simples del corpus d'acord amb el criteri de freqüència. Després aquests termes s'usen en combinació amb heurístiques de caràcter lingüístic i consideracions de freqüència per a extreure altres candidats a terme. Per a fer-ho, té en compte els criteris següents: quan dos termes coneguts apareixen amb freqüència junts, constitueixen un terme complex; quan una paraula apareix amb freqüència amb alguns termes coneguts, la paraula esdevé un terme simple; quan una paraula apareix amb freqüència amb un terme conegut, constitueix un nou terme complex.

Un pas més enllà es fa aprofundint en l'anàlisi lingüística fent servir informació semàntica i contextual. Les estratègies semàntiques serveixen per a afinar els resultats obtinguts amb els mètodes estadístic i lingüístic d'extracció de termes. D'aquestes estratègies n'hi ha bàsicament de dos tipus: d'una banda, les estratègies que fan servir categories semàntiques d'una font lèxica externa al corpus de treball, com WordNet (Miller, 1995), EuroWordNet (Vossen, 1998) o AlethDic (Naulleau, 1998), que organitzen el lèxic a partir del significat de les paraules i que poden integrar-se en una eina d'extracció de candidats de terme; d'altra banda, les que extreuen les categories semàntiques de les paraules del mateix corpus a través d'elements contextuals que fan referència a la combinació sintàctico-semàntica de les paraules, com el model de Fabre (1996).

Concretament, a Naulleau (1998) es fan servir perfils d'usuari per a poder extreure els candidats que satisfacin les necessitats de cada usuari i incorpora informació semàntica.

En el treball de Maynard i Ananiadou (2000a) la informació semàntica que deriva dels tesaurus i la informació lingüística i estadística es combina per a reendreçar els candidats a terme.

Així mateix, NC-value, una mesura heurística complexa, es proposa per a ser combinada amb C-value i el factor de context, la qual té en compte les propietats semàntiques, sintàctiques i estadístiques dels contextos en els quals apareix el candidat a terme. L'ús d'informació extrínseca (com ara el context) és comuna a altres aproximacions. A Trucks (Maynard, 2000; Maynard i Ananiadou, 2000b) es combinen mesures estadístiques amb informació lingüística (morfològica i semàntica) i també informació contextual. A Termine (Frantzi i Ananiadou, 1999; Frantzi *et al.*, 2000; Nenadic *et al.*, 2004; Barrón-Cedeño *et al.*, 2009) es combina l'algorisme híbrid C/NC Value, el qual incorpora amb C-Value l'anàlisi lingüística i estadística i amb NC-Value endreça els candidats a terme segons la informació de context. Al final del procés cada candidat a terme té assignada una puntuació que indica en quin nivell el candidat és qualificat com a terme. L'algorisme C-Value ha estat usat amb èxit en àmbits com ara els receptors nuclears (Ananiadou *et al.*, 2000), les patologies dels ulls (Frantzi *et al.*, 2000), la informàtica (Milios *et al.*, 2003) o les notícies de negocis biomèdics (Zervanou i McNaught, 2004). C-Value és un dels algorismes més usats en els darrers anys per al reconeixement automàtic de termes multiparaula (Wermter i Hahn, 2005b).

El sistema Atract (Mima *et al.*, 2001) (Automatic Term Recognition and Clustering of Terms) incorpora reconeixement automàtic de terminologia, classificació automàtica de termes, recuperació de documents basada en similitud i una base de dades intel·ligent (selecció d'ítems fent servir la base de dades), integrat en un sistema web. La part de reconeixement automàtic de terminologia consta d'una aproximació híbrida que combina coneixement lingüístic (patrons terminològics) i estadístic (freqüència de les ocurrències, llargada de les cadenes, etc.). Aquesta aproximació es fa amb l'algorisme C/NC-Value, que extreu termes multiparaula, identifica termes relacionats (termes aniuats) i selecciona termes a partir de la informació del context. El sistema també permet personalitzar els resultats finals seleccionant un llindar de valor, índex, pes, tria de la categoria gramatical, filtre lingüístic, nombre de paraules incloses en el context, segons les necessitats de l'usuari.

En altres aproximacions, com la de Velardi *et al.* (2001), també s'usa informació extrínseca (com els contextos), en la qual primer s'usa l'anàlisi sintàctica superficial per a seleccionar patrons de candidats a terme i després es fan servir les mesures estadístiques Domain relevance i Domain consensus per a endreçar els termes d'acord amb llurs contextos.

En el treball de Zanzotto (2002) es proposa una ampliació de la definició de terme amb l'objectiu de millorar el procés de reconeixement de termes. Per a l'extracció de termes es fa servir la freqüència com a mesura estadística juntament amb la informació lèxica i sintàctica sobre els contextos en què apareix el terme.

En els treballs de Nakagawa i Mori (2002) i Nakagawa i Mori (2003) es proposa d'extreure unitats terminològiques que corresponen a noms simples i compostos per mitjà de mètodes lingüístics combinats amb mètodes estadístics, els quals assignen una puntuació a cada candidat a terme, que indica la probabilitat que té el candidats de ser un terme. El mètode de puntuació d'un nom simple mesura quants noms compostos diferents contenen el nom simple en qüestió en un corpus o document. Quan tots els candidats tenen una puntuació, són reendreçats en ordre descendent.

Segons Dias *et al.* (2000) i Dias (2003), els mètodes híbrids defineixen coocurrències interessants basant-se en patrons sintàctics i regularitats estadístiques. D'aquesta manera, aquests sistemes redueixen l'àmbit de cerca a grups de paraules que corresponen a priori a patrons sintàctics definits (p. ex. adjectiu-nom, nom-preposició-nom), apliquen la puntuació estadística per a identificar els segments de paraules més interessants i incorporen soroll en els resultats [Ana (Enguehard i Pantera, 1995); Justeson i Katz (1995); Daille (1994, 1995), i Heid (1999)]. Una limitació important d'aquests mètodes és que no tenen en compte el gruix de combinacions multiparaula interessants. I, a més, tenen poca flexibilitat quan els patrons sintàctics s'han de revisar per a ser adaptats a una altra llengua. La proposta que fa Dias (2003) és un sistema híbrid ano-

menat HELAS (Hybrid Extraction of Lexical Associations) el qual extreu unitats multiparaula d'un corpus prèviament anotat morfosintàcticament. A diferència dels mètodes clàssics, identifica automàticament els patrons sintàctics més rellevants. Combina el processament lingüístic amb l'ús de la mesura estadística Mutual Expectation per a obtenir el nivell de cohesió que hi ha en les unitats multiparaula. Aquest sistema introdueix millores respecte els anteriors perquè no necessita tenir definits manualment els patrons sintàctics i pot ser adaptat a qualsevol llengua sense haver-hi de fer adaptacions prèvies.

La proposta de Drouin (2003) amb l'eina TermoStat d'extracció de terminologia es basa en el principi que els termes estan estretament relacionats a l'àrea d'especialitat a la qual pertanyen. Així, l'extracció de termes se centra en la comparació de corpus especialitzats i corpus de caràcter general. Per a fer-ho, el sistema compta amb un corpus d'entrada constituït per dos subcorpus: un d'especialitzat, on es fa la cerca dels termes (corpus d'anàlisi), i un altre de no tècnic (corpus de referència), que serveix per a determinar com és d'estreta la relació d'una paraula respecte al corpus tècnic. L'especificitat d'una paraula respecte del corpus d'anàlisi es basa en el contrast de la freqüència d'aparició en el corpus d'anàlisi respecte al corpus de referència. TermoStat primer etiqueta el corpus i conserva les paraules que tenen la categoria gramatical de nom o adjectiu, per ser les categories més comunes de paraules que tenen un major contingut semàntic en els termes. Són les paraules principals (*headwords*), les quals són el punt de partida d'extracció de candidats i identificació de termes.

A Nenadic *et al.* (2004) es combina el mètode C-Value amb el reconeixement de variació terminològica i s'integra com a part d'un procés d'extracció terminològic. C-Value és una aproximació híbrida que combina patrons de formació terminològica amb mesures estadístiques basades en corpus. La incorporació del tractament de la variació terminològica millora els resultats d'un sistema de reconeixement automàtic de terminologia, en comparació a fer servir únicament el mètode C-Value. Concretament, la precisió augmenta d'un 20% a un 70% en rang de termes més ben si-

tuats, i la cobertura d'un 2% a un 25%. I en un altre treball, Nenadic *et al.* (2005) fa una proposta d'aproximació híbrida que combina patrons i tècniques d'aprenentatge automàtic amb un mecanisme de puntuació estadística per a localitzar similitud lèxica, sintàctica i contextual entre els termes. També combina similituds basades en les correspondències lèxiques internes i diferents tipus de distribucions basades en corpus. Els resultats indiquen que termes similars comparteixen les mateixes associacions. I que els termes que pertanyen semànticament a classes properes també tenen un nivell alt de similitud de context.

RESUM DELS MÈTODES ESTADÍSTICS

- Combinen tècniques lingüístiques i estadístiques.
- Apliquen patrons lingüístics, expressions regulars i mesures d'associació lèxica per a l'extracció de candidats a terme.
- Incorporen informació semàntica de fonts externes al corpus o l'extreuen del mateix corpus amb elements contextuals.
- Tenen poca flexibilitat d'adaptació dels patrons lingüístics d'una llengua a una altra.

En definitiva, els mètodes d'extracció automàtica de terminologia descrits en aquest capítol mostren diverses oportunitats i esculls a l'hora de ser implementats en un procés d'extracció automàtica de terminologia, i també permeten força flexibilitat per a incorporar noves estratègies d'extracció que facin possible una millor identificació automàtica dels termes que són presents en corpus d'especialitat.

3.2 Recapitulació

En aquest capítol hem presentat la disciplina en la qual situem l'extracció automàtica de terminologia i hem descrit els diferents àmbits d'actuació en què té una aplicació directa. La tasca d'extracció automàtica de termes compta amb un paper clau en àmbits relacionats directament amb el tractament lingüístic de corpus especialitzats, com la traducció o la construcció de diccionaris i vocabularis especialitzats, i també en àmbits vinculats amb la cerca, recuperació i gestió d'informació, com la classificació de documents, la indexació textual, la creació de tesaurus, el reconeixement d'entitats amb nom, el resum automàtic o les xarxes semàntiques.

Seguidament hem analitzat amb detall els mètodes que han estat emprats en extracció automàtica de terminologia. Concretament fem una descripció dels mètodes estadístics, lingüístics i híbrids que han estat ideats i implementats amb l'objectiu de poder extreure automàticament d'un corpus els termes que hi són presents.

La revisió feta dels primers sistemes d'extracció de terminologia[1] mostra les mancances que pateixen a l'hora d'extrapolar els resultats a altres àmbits d'especialitat per manca d'indicacions clares i precises de com han estat obtinguts els resultats inicials. A més, els corpus que fan servir són reduïts i molt especialitzats, fet que difícilment permet obtenir resultats adequats amb corpus més grans o menys especialitzats.

Els sistemes que s'han estudiat produeixen molt soroll –sistemes majoritàriament lingüístics– per diferents motius: errors en l'etiquetatge del corpus; dificultat per distingir termes o variants de frases nominals que no són termes; manca de portabilitat a nous àmbits per la dificultat de definir patrons d'anàlisi prou amplis que tinguin una bona precisió; difi-

[1] En el treball de Vivaldi (2001) hi ha un estudi exhaustiu dels primers sistemes d'extracció de terminologia: Ana (Enguehard i Pantera), Acabit (Daille), Clarit (Evans i Zhai), Fastr (Jacquemin), Lexter (Bourigault), Naulleau (Naulleau), Nodalida-95 (Arppe), Termino (David i Plante), Terms (Justeson i Katz) i Trucks (Maynard).

cultat d'avaluació dels resultats sense un patró de referència; avaluacions que varien segons el mètode emprat; sortida de la llista de termes filtrats, que tenen les restriccions del processament i el postprocessament de llistes sobregenerades, o dificultat per identificar els termes a partir de les categories gramaticals (Aubin i Hamon, 2006). Així mateix, silencien un important nombre de termes -–sistemes majoritàriament estadístics. Es produeix un silenci intrínsec al text, és a dir, no detecten les unitats anaforitzades discursivament, i un silenci extrínsec, ja que només detecten uns quants tipus d'unitats especialitzades. A més, la majoria dels sistemes que hi ha actualment s'han dissenyat per al francès o l'anglès i no preveuen que puguin ser emprats per a altres llengües.

Pel que fa al tipus d'unitats que extreuen, aquests sistemes se centren fonamentalment en el sintagma nominal i no en el verb o el sintagma verbal, ja que en els textos especialitzats hi ha molts sintagmes nominals terminològics; ara bé, també seria convenient que tinguessin present els verbs. A més a més, les tècniques d'extracció es fan en cascada i no en paral·lel o combinades, que és el que ofereix més bons resultats.

En aquests sistemes, la informació semàntica es fa servir poc per l'escassetat de recursos de què es disposa. La majoria de sistemes fan servir patrons morfològics per a identificar termes complexos i cobrir la majoria de possibilitats, però no les cobreixen totes i això fa que es produeixi silenci. Sembla que combinar la freqüència d'aparició i els patrons morfològics és adequat per a identificar si una unitat és terminològica. Ara bé, les dades mostren que els resultats ofereixen molt soroll. Els autors de la majoria dels sistemes indiquen que és un error desambiguar la categoria morfològica d'una paraula, però no concreten quin és el grau de la incidència en el resultat.

Finalment, els resultats presenten els candidats a terme aïllats, sense informació complementària per a facilitar-ne la tria o tenir les unitats relacionades per a saber si són adequades o no ho són.

En aquest sentit, és important que els sistemes d'extracció de candidats a terme mostrin la informació de context per a ajudar a triar una unitat terminològica.

L'anàlisi dels diferents mètodes implementats fins al moment, juntament amb una reflexió dels avantatges i les limitacions que proporcionen, ens ha permès plantejar la proposta experimental que descrivim en els capítols 4 i 6.

Capítol 4

EXTRACCIÓ RECURSIVA DE *TOKENS* TERMINOLÒGICS

En el present capítol descrivim un nou mètode recursiu d'extracció automàtica d'unitats terminològiques que hem desenvolupat i que es basa en una estratègia estadística no supervisada d'identificació de termes per mitjà de *tokens* terminològics. Aquest mètode l'hem anomenat *token slot recognition* (mètode TSR) i ha estat implementat en un algorisme que permet fer una selecció recursiva de candidats a terme (CT) d'un corpus prenent com a referència inicial una llista de termes. Amb aquest mètode volem fer una aportació a les limitacions que hem indicat en el capítol 3 respecte les estratègies d'extracció lingüística i estadística de termes.

Seguidament presentem els recursos que formen part de la proposta experimental del mètode TSR (apartat 4.1). A continuació, descrivim en detall el funcionament del procés d'extracció amb el mètode TSR (apartat 4.2). Així mateix, fem una descripció dels resultats obtinguts amb el mètode TSR, els quals són contrastats amb el mètode de freqüència (apartat 4.3). I, finalment, avaluem el rendiment del mètode TSR a partir dels resultats obtinguts amb diferents corpus especialitzats (apartat 4.4).

4.1 Recursos de l'entorn experimental

Per a la implementació del mètode TSR hem seleccionat un conjunt de recursos lingüístics que provenen d'un entorn real de treball terminològic. En aquest sentit, hem volgut evitar la creació de recursos *ad-hoc* per al nostre propòsit experimental. D'una banda, els corpus d'especialitat que hem seleccionat provenen de l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF), el Centre de Terminologia Termcat i el Joint Research Centre de la Unió Europea. D'altra banda, els termes que ens serveixen per a l'avaluació dels resultats obtinguts amb el mètode TSR i que anomenem *termes de referència* han estat seleccionats manualment dels corpus d'especialitat per terminòlegs.

4.1.1 Corpus d'especialitat

Els corpus d'especialitat que hem utilitzat per a analitzar els resultats del mètode TSR han estat seleccionats pel fet de pertànyer a diferents àmbits d'especialitat, tenir un volum de contingut divers, estar disponibles en quatre llengües i també per disposar de tots els termes que hi són presents. La tipologia dels corpus d'especialitat seleccionats permet constatar la rendibilitat que ofereix el mètode TSR en cada un dels casos i veure si és possible l'extrapolació del mètode a altres corpus i llengües.

Dominis d'especialitat i llengües dels corpus

Els dominis d'especialitat dels corpus pertanyen a tres àmbits: l'economia, la medicina i els serveis socials. En l'*àmbit econòmic* s'utilitzen tres corpus provinents de l'Acquis Communautaire, que correspon a un conjunt de textos legislatius de la Unió Europea escrits des de 1950 fins a l'actualitat, els quals són compilats pel Joint Research Centre de la Unió Europea. Aquests corpus els tenim disponibles en tres llengües: anglès, francès i espanyol. El corpus en espanyol té un volum de 13.381 paraules, el corpus en anglès en té 7.863 i el corpus en francès en té 14.188.

Del mateix *àmbit econòmic* també s'empra un corpus compilat per l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF). Aquest corpus és en espanyol i té un volum de 41.385 paraules.

De l'*àmbit mèdic* s'empra un corpus compilat per l'IULA. Aquest corpus és en espanyol i té un volum de 98.532 paraules.

Pel que fa a l'*àmbit dels serveis socials* s'utilitzen un parell de corpus compilats pel Departament de Benestar Social i el Centre de Terminologia Termcat. Aquests corpus els tenim disponibles en dues llengües: en català i en espanyol. El corpus en espanyol té un volum de 26.855 paraules i el corpus en català en té 25.840.

Corpus d'entrenament i corpus de prova

Per tal d'avaluar la rendibilitat del mètode TSR, dividim els corpus d'especialitat en dues parts: el *corpus d'entrenament*, que correspon a la part del corpus d'on seleccionem els termes de referència que serviran de base per a extreure els termes presents en la resta del corpus, que anomenem *corpus de prova*. D'aquesta manera, podem comprovar quina capacitat té el mètode TSR d'extreure termes a partir d'una breu llista de termes de referència que pertanyen al mateix domini d'especialitat i que no són en el corpus de buidatge. El corpus d'entrenament correspon a un trenta per cent del contingut total del corpus i el corpus de prova correspon al setanta per cent restant del corpus.

4.1.2 Termes de referència

Els termes de referència que emprem en la nostra proposta experimental serveixen per a avaluar els resultats obtinguts amb el mètode TSR i així poder-los comparar amb els resultats que s'obtenen amb el mètode de freqüència. Corresponen a tots els termes extrets manualment per terminòlegs dels corpus d'especialitat que acabem de descriure i que hem fet

servir per a implementar el mètode TSR. Aquests termes ens han estat cedits pel Centre de Terminologia Termcat i l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra.[1]

Tal com indiquem en l'apartat 4.2, en la nostra proposta experimental avaluem els resultats obtinguts amb termes bigrams; per aquest motiu, considerem solament els termes de referència bigrams que han estat extrets manualment dels corpus. Així, el nombre de termes bigrams presents en els corpus d'especialitat queda distribuït de la manera següent:

- Corpus d'economia en espanyol de l'IULA: 211 termes.
- Corpus d'economia en espanyol del JRC: 80 termes.
- Corpus d'economia en anglès del JRC: 97 termes.
- Corpus d'economia en francès del JRC: 76 termes.
- Corpus de medicina en espanyol de l'IULA: 307 termes.
- Corpus de serveis socials en espanyol del Termcat: 61 termes.
- Corpus de serveis socials en català del Termcat: 62 termes.

Aquests termes de referència ens serveixen per a definir la rendibilitat que pot oferir el mètode TSR en ser incorporat en un procés d'extracció automàtica de terminologia. Tal com indiquen Vivaldi i Rodríguez (2007), hi ha unes determinades propietats que mostren que un candidat és un terme i corresponen al nivell d'*unithood*, de *termhood* o de pertinença a l'àmbit d'especialitat (ús especialitzat) que tingui; ara bé, aquestes propietats no són gens clares d'identificar. Per aquest motiu, hem considerat oportú d'avaluar la nostra proposta experimental a partir de corpus dels quals disposéssim *a priori* tots els termes, i així poder identificar exactament el nombre de termes que permet recuperar el mètode TSR comparant-lo amb el mètode de freqüència, i extrapolar el mètode a altres àmbits.

[1] Agraïm sincerament la cessió desinteressada del buidatge terminològic dels corpus d'especialitat per a poder-ne fer ús en la present recerca.

4.2 Extracció de termes amb el mètode TSR

La proposta experimental que plantegem en aquest capítol se centra en l'extracció recursiva de candidats a terme de diferents corpus d'especialitat monolingües. La referència que fem a "extracció de candidats a terme" sembla aparentment clara, però a la pràctica es planteja sota diferents punts de vista, pel fet que la tasca d'extracció de terminologia és duta a terme per un ampli grup "d'actors": terminòlegs, especialistes d'àmbits temàtics, experts en extracció d'informació i recuperació d'informació i també traductors. Cada un d'aquests grups té una idea clara sobre la noció de "terme" i "candidat a terme" i, en conseqüència, del resultat final sobre extracció de candidats a terme (Heid, 2006). La present proposta experimental ha estat dissenyada perquè els candidats a terme seleccionats al final del procés d'extracció continguin el màxim nombre de termes representatius de l'àmbit d'especialitat i que puguin ser d'utilitat a un ampli espectre d'actors que intervenen en el treball terminològic. En aquest sentit, es consideren *termes* les unitats lèxiques que tenen una estructura gramatical i un significat específic dins un àmbit d'especialitat, són cohesionades, pròpies d'una determinada llengua i un determinat àmbit en el qual apareixen amb certa freqüència (Huo, 2012). Així mateix, són unitats que no poden ser modificades, pronominalitzades, substituïdes per sinònims, escurçades, que no accepten coordinació ni nominalització dels adjectius (Schmidt, 2001) (sobre la consideració de *terme* vegeu l'apartat 2.1).

El tipus de candidats a terme que extraiem en aquesta proposta experimental són bigrams, tenint en compte que són les unitats que concentren un major nombre de termes i que, de manera general, quan augmenta la longitud dels candidats en nombre d'n-grams disminueix la probabilitat que aquestes unitats siguin terminològicament vàlides. Aquestes consideracions es desprenen de l'anàlisi comparativa que van dur a terme Justeson i Katz (1995) i Nkwenti-Azeh (1994) amb relació al nombre de termes que poden ser recuperats per un sistema d'extracció automàtica i la longitud que tenen aquests termes.

Taula 4.1: Estudis sobre el nombre de termes extrets per longitud.

	1 paraula	2 paraules	3 paraules	4 paraules	5+ paraules
Justeson and Katz	29,5%	54,5%	12,4%	3,9%	0%
Nkwenti-Azeh (1)	9.15%	71.86%	16.93%	2.06%	0%
Nkwenti-Azeh (2)	7.30%	49.02%	32.83%	8.88%	1.97%
Nkwenti-Azeh (3)	30.73%	49.84%	15.13%	3.5%	0.8%

Justeson i Katz van obtenir aquests resultats de les dades extretes de quatre diccionaris anglesos (fibra òptica, medicina, física i matemàtiques i psicologia), les quals van ser confirmades amb proves fetes en diferents corpus. I Nkwenti-Azeh va treballar amb dades extretes de diferents fonts d'informació en anglès, com ara corpus (1), bases de dades terminològiques (2) i diccionaris tècnics (3).

El mètode TSR es basa en una estratègia estadística no supervisada amb l'objectiu de poder ser implementat en projectes de buidatge terminològic en els quals intervinguin diferents llengües i en els quals s'hagin de seleccionar automàticament els candidats a terme que siguin propis d'un àmbit d'especialitat. En aquest sentit, l'aplicació d'estratègies estadístiques en mètodes d'extracció automàtica de termes ofereix major flexibilitat, és a dir, permet adaptar el mètode a noves llengües, a diferència de les estratègies lingüístiques, en les quals s'han d'adaptar els patrons lingüístics a les diferents llengües de treball.

Així mateix, la recursivitat del mètode TSR se centra en l'aprofitament dels candidats a terme que estan constituïts per *tokens* terminològics en el procés iteratiu d'extracció de termes, amb la finalitat d'obtenir el màxim nombre d'unitats terminològiques del corpus especialitzat del qual es duu a terme el buidatge. Si considerem que un *token* és una seqüència contínua de paraules, podem dir que un *token* terminològic és una seqüència contínua de paraules presents específicament en els termes. Un candidat

a terme pot estar constituït per *n tokens* i els termes per *n tokens* terminològics. El mètode TSR identifica els *tokens* terminològics presents en termes propis d'un àmbit d'especialitat i selecciona els corresponents candidats a terme a partir de la presència o absència de *tokens* terminològics en aquests candidats.

L'estratègia de la recursivitat en extracció de candidats a terme ha estat aplicada amb bons resultats en diverses ocasions. Concretament, en una aproximació híbrida (lingüística i estadística) d'extracció de sintagmes nominals (CLARIT), els quals són analitzats en un procés recursiu per tal de localitzar àtoms lèxics que són els que mantenen relacions més estretes i segures (Evans i Zhai, 1996). Així mateix, ha estat aplicada en la constitució de corpus especialitzats presents a la xarxa i l'extracció dels termes corresponents (Baroni i Bernardini, 2004). Ha estat incorporada en mètodes lingüístics d'extracció de termes, concretament en la fase d'anàlisi fragmental (*chunking*) de corpus especialitzats (Heid, 2006). I de la mateixa manera s'ha posat en pràctica en una aproximació iterativa basada en el mètode de remostreig (*bootstrapping*) aplicada a l'extracció de termes propis d'un àmbit d'especialitat en textos sense anotar. En aquest cas, les cadenes formades per components amb una major probabilitat de ser termes específics d'un àmbit són identificades com a candidats a terme (Zhang *et al.*, 2012).

4.2.1 Descripció del mètode TSR

En el procés d'extracció de termes del mètode TSR, l'algorisme assigna a cada un dels candidats la condició de terme si aquests són formats per *tokens* terminològics presents en els termes d'un àmbit d'especialitat. En aquest sentit, dos candidats a terme com ara "asma aguda" o "afectación bronquial" de l'àmbit de medicina contenen dos *token slots* dels quals el *token slot 1* és ocupat per "asma" i "afectación" i el *token slot 2* per "aguda" i "bronquial". L'algorisme seleccionarà aquests candidats a terme si un *token slot* o més d'un pot ser ocupat per *tokens* terminològics de la llista de termes, com ara "asma bronquial", "bronquitis aguda" i "afecta-

ción nutricional" (Wermter i Hahn, 2005a). Tal com s'observa en la taula 4.2, l'algorisme seleccionarà com a termes ambdós candidats, els quals al final del procés de revisió manual seran considerats termes de l'àmbit de medicina. Ara bé, en la mateixa taula es mostra com altres candidats que també han estat seleccionats per l'algorisme com a possibles termes, finalment no són recollits com a tals ("inflamación bronquial", "asma severa", "mucosa bronquial", "asma nocturna").

Taula 4.2: Selecció de candidats a terme.

Termes referència	**CT posició 1**	**CT posició 2**	**Termes posició 1**	**Termes posició 2**
asma bronquial	asma aguda	inflamación bronquial	asma aguda	
bronquitis aguda	asma severa	hiperreactividad bronquial		hiperreactividad bronquial
afectación nutricional	asma episódica	mucosa bronquial	asma episódica	
	asma nocturna	afectación bronquial		afectación bronquial
	bronquitis asmatiforme		bronquitis asmatiforme	
	afectación bronquial		afectación bronquial	

Aquest mètode requereix un conjunt reduït de termes inicial per a poder començar el procés de selecció de candidats. En cas que no hi hagi cap llista de termes disponible, es pot iniciar el procés fent la selecció manual dels candidats. Els termes que hagin estat seleccionats manualment seran utilitzats per l'algorisme com a termes de referència per a preparar el procés recursiu d'extracció de candidats a terme.

4.2.2 Procés d'extracció automàtica de termes

El procés d'extracció automàtica de termes que duu a terme l'algorisme consta específicament de quatre passos, que descrivim a continuació.

Pas 1

Primerament s'extreuen els candidats a terme dels corpus d'especialitat, els quals serviran de base perquè el mètode TSR pugui identificar quins són termes a partir del filtratge recursiu amb *tokens* terminològics. L'extracció dels candidats a terme es duu a terme amb l'eina d'anàlisi estadística Ngram Statistics Package (Text-NSP). Aquesta eina fou creada i desenvolupada en Perl per Ted Pedersen, Satanjeev Banerjee, Amruta Purandare, Bridget Thomson-McInnes i Saiyam Kohli l'any 2001, s'actualitza periòdicament, és de codi lliure i es distribueix amb una llicència pública general de GNU. L'eina Text-NSP genera la llista de candidats a terme d'un corpus, filtrats per paraules buides i endreçats per freqüència, a partir d'un fitxer de text pla.

Pas 2

Seguidament l'algorisme contrasta la llista de candidats a terme extrets en el pas 1 amb els termes de referència seleccionats manualment per terminòlegs dels corpus d'especialitat, a fi d'identificar quins candidats estan formats per *tokens* terminològics en posició inicial, en posició final o en totes dues posicions del bigram. Per *token* terminològic considerem una seqüència contínua de caràcters present específicament en els termes. Així, en aquest punt del processament es duu a terme el filtratge de candidats els *tokens* dels quals siguin coincidents amb els *tokens* terminològics presents en els termes. Com a resultat del filtratge s'obté una llista de candidats a terme que estan constituïts parcialment o totalment per *tokens* terminològics. Aquest resultat serveix de base per a dur a terme el processament del pas 3.

Pas 3

A continuació els candidats a terme que estan constituïts per *tokens* terminològics i que són termes passen a formar part de la llista de termes de referència per a completar-la. Els nous termes que s'han extret del corpus i que amplien la llista de termes de referència permeten identificar un major nombre de candidats susceptibles de ser considerats termes en la següent iteració del procés de filtratge. En aquest punt del procés, l'algorisme torna a contrastar els candidats a terme extrets inicialment dels corpus amb la llista ampliada de termes de referència. Així, es tornen a filtrar els candidats que estan formats per *tokens* terminològics. Del resultat d'aquesta segona extracció, els candidats que són termes s'incorporen a la llista de termes de referència. El procés d'incorporar els candidats que són termes a la llista de termes de referència és recursiu, es produeix en cada iteració, i es repeteix fins a arribar a la saturació dels resultats, és a dir, fins que no es poden identificar més candidats que tinguin caràcter terminològic.

Pas 4

En el darrer pas del procés de filtratge de candidats es duu a terme la revisió manual final dels resultats obtinguts, a fi de validar els termes que han estat extrets del corpus. Tenint en compte que en aquest cas disposem de la llista completa de termes que formen part dels corpus, solament s'han de validar manualment els candidats que han estat incorporats automàticament com a termes pel fet d'estar formats per *tokens* terminològics per tal de comprovar la seva adequació terminològica.

Seguidament descrivim el pseudoalgorisme corresponent al mètode TSR i presentem de manera gràfica en la figura 4.1 el procés d'extracció d'automàtica de termes.

ALGORISME TOKEN SLOT EXTRACTION (TSR)

Entrada: corpus especialitzat
Sortida: llista de termes endreçats per freqüència

Mètode:
Extracció de bigrams del corpus especialitzat {
Entrada: corpus
Sortida: bigrams
Programa: Text-NSP }
Filtratge (bigrams per paraules buides)
Ordre (bigrams per frequència, descendent)
Sortida (llista de bigrams endreçats per freqüència)

Filtratge de bigrams per tokens terminològics {
Entrada: termes de referència
Sortida: candidats a terme
Programa: token slot extraction
Ordre (candidats per freqüència, descendent)
Selecció dels candidats que són termes
Actualització llista de termes referència (nous termes extrets)
} n vegades

Sortida (llista de termes filtrats per tokens terminològics)
Revisió manual final dels resultats

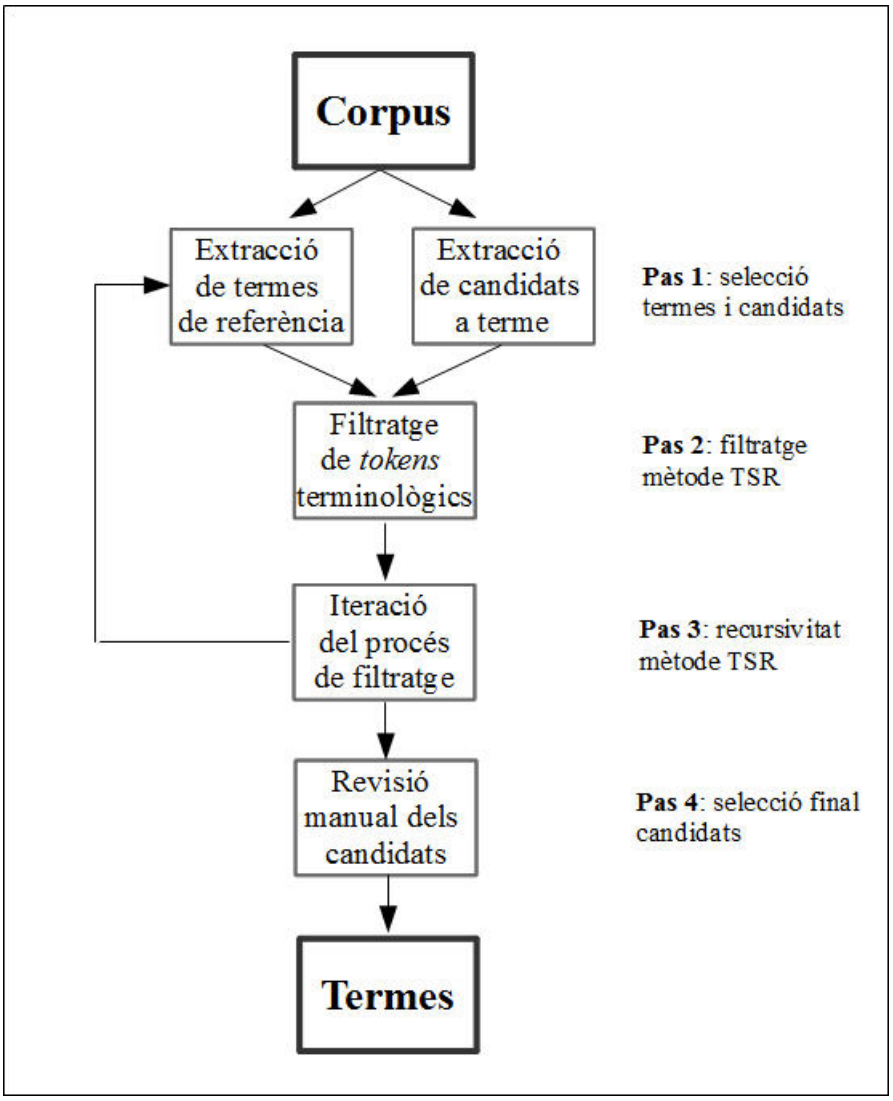


Figura 4.1: Extracció de terminologia amb el mètode TSR.

4.3 Descripció dels resultats

La implementació del mètode TSR es duu a terme en un conjunt de corpus especialitzats de diversos àmbits d'especialitat amb un volum de contingut diferenciat (apartat 4.1.1). Així mateix, la rendibilitat que assoleix aquest mètode és contrastada amb el mètode de freqüència, una mesura utilitzada àmpliament en extracció de terminologia com a punt de referència per al contrast de resultats i que té en compte els candidats que apareixen amb major freqüència en un corpus (Evert i Krenn, 2001). Seguidament presentem el detall dels resultats d'extracció de termes obtinguts amb el mètode TSR i el mètode freqüència tenint en compte, d'una banda, el nombre de termes que extreu cada mètode considerant diferents mostres de resultats, i, d'altra banda, la quantitat de termes que extreu en cada una de les iteracions el mètode TSR.

Amb relació als resultats que s'obtenen comparant diferents mostres de resultats, convé observar que hem mirat d'agafar mostres de candidats que fossin similars en nombre i quantitat en tots els corpus. Ara bé, tenint en compte que treballem amb corpus de diferent volum, la quantitat total de mostres que presentem varia segons el volum de candidats que hem extret de cada corpus. Així, per corpus indiquem el nombre de candidats que componen les mostres que hem seleccionat i, a sota, anotem el nombre de termes que identifica el mètode TSR i el mètode freqüència per mostra. Aquesta estructura la repetim per cada corpus. En la taula 4.3 recollim la distribució dels resultats.

Quant als resultats que s'obtenen a partir de les iteracions del mètode TSR, cal dir que en aquest cas recollim els processos de filtratge per *tokens* terminològics que duu a terme el mètode TSR. En aquest sentit, comparem els resultats obtinguts en les iteracions del mètode TSR amb els resultats equivalents que assoleix el mètode freqüència. Concretament, recollim el nombre de termes identificat amb el mètode freqüència segons el nombre de candidats que hi ha en cada iteració. El detall de resultats per nombre d'iteracions és recollit en les taules 4.6 i 4.7.

Resultats per nombre de candidats a terme

A continuació recollim el nombre de termes que permeten extreure el mètode TSR i el mètode freqüència en les diferents mostres de candidats a terme analitzades.

Taula 4.3: Nombre de termes extrets per mètode.

	Economia espanyol JRC					**Economia anglès JRC**				
Candidats	10	50	100	150		10	50	100	200	350
Termes TSR	9	30	43	55		8	28	41	54	81
Termes freq.	6	25	35	43		5	23	33	46	58
	Economia francès JRC					**Economia espanyol IULA**				
Candidats	10	50	90			10	50	100	250	400
Termes TSR	8	21	36			10	42	65	107	145
Termes freq.	3	14	19			6	25	50	84	114
	Serveis socials cat. Termcat					**Serveis socials esp. Termcat**				
Candidats	10	50	80			10	50	80		
Termes TSR	8	17	27			7	22	34		
Termes freq.	0	8	12			2	10	14		
	Medicina espanyol IULA									
Candidats	10	50	100	200	400	600	750			
Termes TSR	8	26	40	68	103	131	155			
Termes freq.	4	16	25	35	62	74	82			

Els resultats indiquen que en les diferents mostres analitzades el mètode TSR extreu un major nombre de termes que el mètode freqüència. Així, pel que fa al primer corpus d'economia en espanyol del JRC observem que en els primers 10 candidats a terme amb el mètode TSR s'obtenen 9 termes i amb el mètode freqüència se n'obtenen solament 6; en la mostra de 50 candidats el mètode TSR identifica 30 termes i el de freqüència 25 termes; en la posició 100 amb el mètode TSR s'obtenen 43 termes i

amb freqüència 35 termes, i en la mostra de 150 candidats amb el mètode TSR es recuperen 55 termes i amb el mètode freqüència únicament 43. Aquest patró de resultats es reprodueix en tots els corpus i, en conseqüència, constatem que el mètode TSR permet extreure un nombre més elevat de termes que no pas el mètode de freqüència en les diferents mostres de candidats a terme avaluades.

Per tal d'observar el tipus de candidats i termes que s'extreuen amb els mètodes TSR i freqüència, en les taules 4.4 i 4.5 recollim una mostra dels primers resultats obtinguts amb tots dos mètodes i indiquem quin tipus d'unitat identifica cada mètode, terme (T) o candidat a terme (CT). Aquests exemples constaten que els candidats que filtra el mètode TSR tenen major caràcter terminològic que no pas els candidats extrets amb el mètode freqüència, pel fet d'estar constituïts per *tokens* terminològics. Aquesta diferència permet tenir agrupades en els resultats d'extracció les unitats terminològiques extretes dels corpus i, en conseqüència, en facilita la localització en la darrera fase de revisió manual dels resultats.

Taula 4.4: Mostra de candidats i termes amb TSR i freqüència (I).

Economia anglès JRC			
Mètode TSR		Mètode freqüència	
euro area	T	Member States	CT
internal market	T	euro area	T
Lisbon strategy	CT	growth potential	CT
productivity growth	T	Lisbon strategy	CT
labour market	T	European Council	CT
macroeconomic policies	T	internal market	T
employment policies	T	labour market	T
economic growth	T	macroeconomic policies	T
economic recovery	T	productivity growth	T
European economy	CT	structural reforms	T

Taula 4.5: Mostra de candidats i termes amb TSR i freqüència (II).

Economia francès JRC			
Mètode TSR		Mètode freqüència	
politiques macroéconomiques	T	États membres	CT
zone euro	T	lignes directrices	CT
politiques économiques	T	zone euro	T
croissance économique	T	politiques macroéconomiques	T
reprise économique	T	Union européenne	CT
base industrielle	CT	politiques économiques	T
capital humain	T	Ligne directrice	CT
demande intérieure	T	Voir également	CT
cohésion sociale	T	programmes nationaux	CT
croissance durable	CT	ligne directrice	CT
Serveis socials espanyol Termcat			
Mètode TSR		Mètode freqüència	
iniciativa social	T	servicios sociales	T
atención social	T	Servicios Sociales	T
discapacidad intelectual	T	presente ley	CT
violencia machista	T	entes locales	CT
financiación pública	CT	Consejo General	CT
asistente personal	T	sistema público	CT
innovación tecnológica	CT	departamento competente	CT
titularidad pública	T	iniciativa social	T
atención domiciliaria	T	atención social	T
iniciativa mercantil	CT	Atención Pública	CT
Medicina espanyol IULA			
Mètode TSR		Mètode freqüència	
asma bronquial	T	asma bronquial	T
hiperreactividad bronquial	T	hiperreactividad bronquial	T
función pulmonar	T	Asma bronquial	T
obstrucción bronquial	T	función pulmonar	T
grupo control	CT	grupo control	CT
reactividad bronquial	T	obstrucción bronquial	T
sexo masculino	T	efectos colaterales	CT
flujo espiratorio	T	doble ciego	CT
Asma bronquial	T	provocación bronquial	CT
asma infantil	T	sexo masculino	T

Resultats per iteracions

El mètode TSR extreu els termes a partir d'un procés iteratiu en el qual filtra els candidats que estan formats per *tokens* terminològics. Els resultats obtinguts en cada una de les iteracions són recollits en les taules 4.6 i 4.7. Així mateix, contrastem aquests resultats amb el nombre de termes que el mètode freqüència permet recuperar en cada iteració.

Taula 4.6: Nombre de termes extrets per iteració (I).

Economia espanyol JRC				
Iteració	Candidats	Candidats nous	Termes TSR	Termes freqüència
1	138	138	54	40
2	157	19	56	43
Economia anglès JRC				
Iteració	Candidats	Candidats nous	Termes TSR	Termes freqüència
1	302	302	77	55
2	358	56	82	58
3	365	7	83	59
Economia francès JRC				
Iteració	Candidats	Candidats nous	Termes TSR	Termes freqüència
1	90	90	36	19
2	97	7	36	19

Taula 4.7: Nombre de termes extrets per iteració (II).

Economia espanyol IULA				
Iteració	Candidats	Candidats nous	Termes TSR	Termes freqüència
1	283	283	112	88
2	384	101	140	112
3	399	15	146	114
4	413	14	147	115
5	414	1	148	116
Serveis socials espanyol Termcat				
Iteració	Candidats	Candidats nous	Termes TSR	Termes freqüència
1	77	77	33	14
2	84	7	35	15
3	86	2	35	17
Serveis socials català Termcat				
Iteració	Candidats	Candidats nous	Termes TSR	Termes freqüència
1	76	76	26	12
2	80	4	27	12
Medicina espanyol IULA				
Iteració	Candidats	Candidats nous	Termes TSR	Termes freqüència
1	722	722	146	81
2	790	68	157	86
3	794	4	159	86

Els resultats per iteracions mostren el procés d'extracció de termes amb el mètode TSR. En la primera iteració s'obté un nombre elevat de candidats a terme que estan formats per *tokens* terminològics en la primera posició de l'n-gram, en la segona posició o en totes dues. La identificació de *tokens* terminològics es duu a terme comparant els *tokens* presents en els candidats i en els termes de referència. Si hi ha coincidència, el candidat és seleccionat com a unitat terminològica i passa a formar part de la llista de termes de referència. Si observem els resultats corresponents al corpus de medicina en espanyol de l'IULA, constatem que en la primera iteració l'algorisme ha seleccionat 722 candidats dels quals 146 són termes. Així, aquests 146 termes s'afegeixen a la llista de termes de referència per a poder dur a terme la segona iteració del procés d'extracció. En aquest punt del procés, el nombre de candidats nous disminueix considerablement i, alhora, és possible identificar més termes. Observem que en el corpus de medicina la segona iteració recupera 68 candidats nous i permet extreure 11 termes més. En cas de no disposar de la llista de termes presents en el corpus, aquest procés iteratiu permet anar identificant nous termes sense haver d'invertir massa temps en la revisió manual dels candidats a partir de la segona iteració. L'extracció per iteració del mètode TSR es duu a terme fins a arribar a la saturació dels resultats obtinguts, és a dir, fins que ja no es puguin filtrar més termes. Juntament amb el nombre de termes que s'extreuen en cada iteració amb el mètode TSR, aportem els resultats que s'obtenen amb el mètode freqüència. Així, indiquem el nombre de termes que situa el mètode freqüència en el volum de candidats extrets en cada iteració. D'aquesta manera podem observar, a partir d'un mateix volum de candidats, quin mètode te més capacitat d'identificar els termes presents en un corpus. Concretament, veiem que en la primera iteració del corpus de medicina en espanyol de l'IULA el mètode TSR extreu 146 termes i el mètode freqüència únicament 81; en la segona iteració el mètode TSR acumula 157 termes i el mètode freqüència 86, i en la tercera iteració el mètode TSR acaba acumulant 159, en contraposició amb el mètode freqüència, que acaba identificant 86 termes. Així, doncs, els resultats per iteració constaten que el mètode TSR té una major capacitat d'extreure termes que no pas el mètode freqüència.

L'anàlisi detallada dels resultats obtinguts també mostra la capacitat que té el mètode TSR per a identificar nous termes, ja sigui per mitjà de la combinació dels *tokens* presents en els termes de referència o bé a partir d'un dels *tokens* presents en aquests termes. El filtratge de candidats per mitjà de *tokens* terminològics recupera els termes de referència de què disposem per cada corpus i fa possible identificar noves unitats terminològiques que inicialment no tenim disponibles com a termes, però que poden ser considerades com a tals, tenint en compte que part dels seus components o ambdós són presents en els termes de referència. Aquesta capacitat que té el mètode TSR és rellevant sobretot per a poder identificar nous termes en corpus dels quals no disposem d'una llista prèvia de termes de referència i situar-los en les posicions inicials de la llista de resultats.

Amb l'objectiu d'observar amb detall com s'han identificat aquestes noves unitats, hem seleccionat quatre dels corpus que hem emprat en la present recerca i que pertanyen a diferents àmbits d'especialitat, són de diferents llengües i tenen un volum divers. Aixi, en les taules 4.8, 4.9, 4.10 i 4.11 recollim una mostra de quins són els termes de referència a partir dels quals s'han localitzat nous termes. En cada cas hem marcat en cursiva el *token* terminològic que ha servit de base per a extreure cada nou terme, els quals poden ser agrupats en una d'aquestes tres categories.

a) Nous termes formats per una combinació de *tokens* terminològics

social protection, *market reforms*, *cadre multilatéral*, *environnement économique*, *mobilité européenne*, *participación económica*, *participación comunitaria*, *administración pública*, *unidad familiar*, *provocación bronquial*, *efecto broncodilatador*, *acción broncodilatadora*, *crisis aguda*, *asma severa*, *membrana celular*

b) Nous termes constituïts per un sol *token* terminològic situat en la posició inicial del terme de referència

public procurement, *sustainable development*, *coûts environnementaux*, *coûts administratifs*, *discapacidad visual*, *centro proveedor*, *capacidad vital*, *estudio multicéntrico*

c) Nous termes constituïts per un sol *token* terminològic situat en la posició final del terme de referència

potential growth, *product markets*, *insertion sociale*, *infrastructures nationales*, *estabilidad laboral*, *red pública*, *población infantil*

Cal dir que la identificació d'aquestes noves unitats terminològiques és clau per a poder anar ampliant recursivament la llista de termes, i aixi poder completar els resultats finals d'extracció automàtica de termes d'un corpus especialitzat. En aquest sentit, el mètode TSR té major capacitat d'identificar nous termes com més volum tingui el corpus, perquè disposa d'una major presència de *tokens* terminològics.

Taula 4.8: Identificació de nous termes (I).

Economia anglès JRC	
Termes de referència	Termes nous
public finances *public* sector	public procurement
economic *growth* employment *growth*	potential growth
sustainable growth	sustainable development
competitive economy *competitive* markets	competitive advantages
labour market *labour* supply	labour cost
social cohesion environmental *protection*	social protection
employment creation *employment* growth	employment security
economic conditions *economic* developments	economic system
competitive *markets* liberalised *markets*	product markets
economic conditions *economic* developments	economic incentives
public finances *public* sector	public expenditure
policy objectives	policy coordination
competitive economy *competitive* markets	competitive environment
market regulation microeconomic *reforms*	market reforms
competitive *markets* labour *markets*	financial markets

Taula 4.9: Identificació de nous termes (II).

Economia francès JRC	
Termes de referència	Termes nous
coûts sociaux *coûts* économiques	coûts environnementaux
cadre macroéconomique surveillance *multilatérale*	cadre multilatéral
coûts sociaux *coûts* économiques	coûts administratifs
environnement macroéconomique activité *économique* avance *économique*	environnement économique
cohésion *sociale* durabilité *sociale*	insertion sociale
politiques *nationales*	infrastructures nationales
politiques *macroéconomiques*	directrices macroéconomiques
mobilité géographique *mobilité* intersectorielle excellence *européenne* économie *européenne*	mobilité européenne
mesures correctrices *mesures* incitatives	mesures protectionnistes
secteurs économiques	secteurs industriels
cadre macroéconomique	cadre communautaire
capital humain	capital risque

Taula 4.10: Identificació de nous termes (III).

Serveis socials espanyol Termcat	
Termes de referència	Termes nous
inserción *laboral* integración *laboral*	estabilidad laboral
responsabilidad *pública* titularidad *pública*	red pública
participación ciudadana prestación *económica*	participación económica
participación ciudadana dimensión *comunitaria*	participación comunitaria
administración relacional titularidad *pública*	administración pública
discapacidad física *discapacidad* intelectual	discapacidad visual
centro abierto *centro* ocupacional	centro proveedor
asistente *personal* autonomía *personal*	renta personal
adaptación *social* aislamiento *social*	fragmentación social
protección *jurídica*	personalidad jurídica
unidad convivencial desestructuración *familiar*	unidad familiar
atención básica *atención* diurna	atención psicológica
integración familiar *integración* laboral	integración sociolaboral

Taula 4.11: Identificació de nous termes (IV).

Medicina espanyol IULA	
Termes de referència	Termes nous
provocación nasal afectación *bronquial* asma *bronquial*	provocación bronquial
capacidad cardíaca *capacidad* contráctil	capacidad vital
adrenoleucodistrofia *infantil* asma *infantil* salud *infantil*	población infantil
efecto clínico *efecto* colateral medicamento *broncodilatador*	efecto broncodilatador
acción antiinflamatoria actividad *broncodilatadora* droga *broncodilatadora*	acción broncodilatadora
estudio anatomopatológico *estudio* angiográfico	estudio multicéntrico
crisis asmática asma *aguda* bronquitis *aguda*	crisis aguda
asma aguda *asma* alérgica anemia *severa* hipocapnia *severa*	asma severa
membrana plasmática *membrana* respiratoria infiltración *celular*	membrana celular

4.4 Avaluació del mètode TSR

En situacions en les quals s'ha de resoldre un problema binari, els resultats que s'obtenen són classificats com a positius o negatius. Les decisions que pren l'eina que classifica aquests resultats queden representades en una estructura coneguda com a matriu de confusió o taula de contingència. La matriu de confusió consta de quatre categories: els veritables positius (VP) corresponen a candidats correctament identificats com a termes. Els falsos positius (FP) són candidats que han estat identificats erròniament com a termes. Els veritables negatius (VN) corresponen a candidats que no són termes i que han estat classificats correctament com a tals. I, finalment, els falsos negatius (FN) fan referència a candidats que són termes i que han estat incorrectament classificats com a no termes (taula 4.12) (Davis i Goadrich, 2006).

Taula 4.12: Matriu de confusió.

	Positius reals	**Negatius reals**
Positius predits	VP	FP
Negatius predits	FN	VN

Donada la matriu de confusió, les mètriques usades en cada espai de la matriu són definides de la manera següent:

Taula 4.13: Definició de mètriques.

Precisió	$= \frac{TP}{TP+FP}$
Cobertura	$= \frac{TP}{TP+FN}$
Índex de veritables positius	$= \frac{TP}{TP+FN}$
Índex de falsos positius	$= \frac{FP}{FP+TN}$

Les mètriques d'avaluació estàndard que són emprades en extracció automàtica de terminologia corresponen a les mètriques de precisió i de cobertura (tal com s'observa en Krauthammer i Nenadic (2004), Vivaldi i Rodríguez (2007), Ha (2007), Korkontzelos *et al.* (2008), Huo (2012), Ittoo i Bouma (2013) o Gojun *et al.* (2012)). En aquest sentit, la precisió mesura l'adequació de les unitats lèxiques proposades com a termes, que correspon a la ràtio entre el nombre de termes correctes (*veritables positius*) i el nombre total d'unitats proposades (*veritables positius* i *falsos positius*). I la cobertura assenyala el nivell amb què els termes són identificats en els corpus, que correspon a la ràtio entre el nombre de termes identificats correctament (*veritables positius*) i el nombre total de termes (*veritables positius* i *falsos negatius*). En l'aplicació d'aquestes dues mètriques en processos d'avaluació sovint passa que si se n'ajusta una l'altra queda desajustada, motiu pel qual s'opta per obtenir sempre una bona cobertura malgrat perdre precisió, la qual queda corregida amb un postprocessament manual dels resultats (Alegria *et al.*, 1999).

L'avaluació del mètode TSR es duu a terme aplicant les mètriques de precisió i cobertura i prenent com a referència els termes presents en cada un dels corpus especialitzats. Per a fer-ho, primerament es preparen els corpus especialitzats dividint-los en dues parts: una part d'entrenament, que correspon a un trenta per cent del contingut del corpus, i una part de prova, que correspon al setanta per cent del total. Havent dividit els corpus, els termes que són presents en el corpus d'entrenament el mètode TSR els utilitza per a seleccionar els candidats a terme del corpus de prova. Al final del procés d'extracció, els termes obtinguts del corpus de prova permeten identificar el nivell de cobertura i precisió que assoleix el mètode. En la taula 4.14 hi ha recollit el detall de com queden distribuïts en volum els corpus d'entrenament i prova, i en nombre els termes.

Taula 4.14: Distribució dels corpus d'avaluació.

Corpus especialitzats	**Procedència corpus**	**Corpus entren.**	**Termes entren.**	**Corpus prova**	**Termes prova**
Economia (es)	JRC	4.014	77	9.367	80
Economia (an)	JRC	2.358	74	5.505	97
Economia (fr)	JRC	4.256	50	9.932	76
Economia (es)	IULA	12.415	109	28.970	211
Serveis socials (ca)	Termcat	7.752	34	18.088	62
Serveis socials (es)	Termcat	8.056	31	18.799	61
Medicina (es)	IULA	29.559	181	68.973	307

Seguidament, els candidats a terme extrets del corpus de prova són seleccionats amb el mètode TSR a partir dels *tokens* terminològics presents en els termes d'entrenament. En aquest procés es duu a terme el contrast de cada un dels *tokens* dels candidats a terme amb els *tokens* terminològics dels termes d'entrenament i s'obté una llista de candidats a terme seleccionats a partir dels *tokens* terminològics. En aquest estadi, l'algorisme assigna a cada un dels candidats l'estat de terme si un dels *tokens* dels quals es compon el candidat coincideix amb els *tokens* terminològics dels termes d'entrenament. A partir d'aquí, la selecció recursiva de candidats s'activa si els candidats a terme seleccionats contenen *tokens* terminològics. La recursivitat del mètode TSR finalitza quan no hi ha cap *token* dels candidats a terme que sigui coincident amb els *tokens* terminològics dels termes d'entrenament. A continuació es mostren els resultats d'extracció obtinguts de cada corpus especialitzat amb el mètode TSR i el mètode de freqüència. Concretament, els resultats mostren el nombre de termes extrets en cada iteració del procés de selecció de candidats i el corresponent valor de precisió i cobertura. En els resultats també s'hi inclou la mesura-F, que és un valor únic que combina precisió i cobertura.

$$Mesura - F = 2 \times \frac{precisió \times cobertura}{precisió + cobertura}$$

Comparació de resultats

A continuació mostrem els resultats de l'avaluació obtinguts amb el mètode TSR, els quals són contrastats amb els del mètode de freqüència.

Taula 4.15: Avaluació de resultats dels corpus d'economia (I).

Corpus economia espanyol JRC						
Mètode TSR						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	138	138	54	40,00	67,50	50,23
2	157	19	56	36,13	70,00	47,66
Mètode freqüència						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	138	138	40	29,63	50,00	37,21
2	157	19	43	27,74	53,75	36,60
Corpus economia anglès JRC						
Mètode TSR						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	302	302	77	25,25	79,38	38,31
2	358	56	82	23,10	84,54	36,28
3	365	7	83	22,74	85,57	35,93
Mètode freqüència						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	302	302	55	18,03	56,70	27,36
2	358	56	58	16,34	59,79	25,66
3	365	7	59	16,16	60,82	25,54

Taula 4.16: Avaluació de resultats dels corpus d'economia (II).

Corpus economia francès JRC						
Mètode TSR						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	90	90	36	42,35	47,37	44,72
2	97	7	36	37,89	47,37	42,11
Mètode freqüència						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	90	90	19	21,11	25,00	22,89
2	97	7	19	20,00	25,00	22,22
Corpus economia espanyol IULA						
Mètode TSR						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	283	283	112	39,30	53,08	45,16
2	384	101	140	36,36	66,35	46,98
3	399	15	146	36,50	69,19	47,79
4	413	14	147	35,85	69,67	47,34
5	414	1	148	35,66	70,14	47,28
Mètode freqüència						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	283	283	88	30,88	41,71	35,48
2	384	101	112	29,09	53,08	37,58
3	399	15	114	28,50	54,03	37,32
4	413	14	115	28,05	54,50	37,04
5	414	1	116	27,95	54,98	37,06

Taula 4.17: Avaluació de resultats dels corpus de serveis socials.

Corpus serveis socials català Termcat						
Mètode TSR						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	76	76	26	34,21	41,94	37,68
2	80	4	27	33,75	43,55	38,03
Mètode freqüència						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	76	76	12	15,79	19,35	17,39
2	80	4	12	15,00	19,35	16,90
Corpus serveis socials espanyol Termcat						
Mètode TSR						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	77	77	33	42,86	54,10	47,83
2	84	7	35	41,67	57,38	48,28
3	86	2	35	40,70	57,38	47,62
Mètode freqüència						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	77	77	14	18,18	22,95	20,29
2	84	7	15	17,86	24,59	20,69
3	86	2	17	19,77	27,87	23,13

Taula 4.18: Avaluació de resultats del corpus de medicina.

Corpus medicina espanyol IULA						
Mètode TSR						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	722	722	146	20,28	47,56	28,43
2	790	68	157	20,00	51,14	28,75
3	794	4	159	20,13	51,79	28,99
Mètode freqüència						
Iteració	CT	CT nous	Termes	Precisió	Cobertura	Mesura-F
1	722	722	81	11,25	26,38	15,77
2	790	68	86	10,96	28,01	15,75
3	794	4	86	10,89	28,01	15,68

Els corpus d'economia en espanyol de l'IULA i el JRC i en anglès del JRC assoleixen els resultats de cobertura més alts, els quals se situen entre el 70 i el 85,5 per cent d'encert. Així mateix, s'observa que a major nombre de termes d'entrenament major és el nombre de termes que s'extreu del corpus de prova. A més, l'anàlisi de corpus de diferents àmbits d'especialitat i volum mostra que la cobertura dels resultats no està relacionada amb la mida del corpus, i és que la cobertura més alta de resultats s'esdevé en un corpus gran (economia espanyol IULA), en un corpus mitjà (economia espanyol JRC) i en un corpus petit (economia anglès JRC).

Si es fa una comparació dels resultats obtinguts amb el mètode TSR i amb el mètode de freqüència, el qual endreça els candidats a terme per freqüència d'aparició en el corpus, s'observa que la cobertura i la precisió del mètode TSR són superiors a les que s'obté amb la freqüència. Concretament, la diferència de cobertura se situa entre un 15 i un 38 per cent i la diferència de precisió entre un 8 i un 21 per cent. Seguidament mostrem gràficament els resultats obtinguts amb tots dos mètodes per corpus.

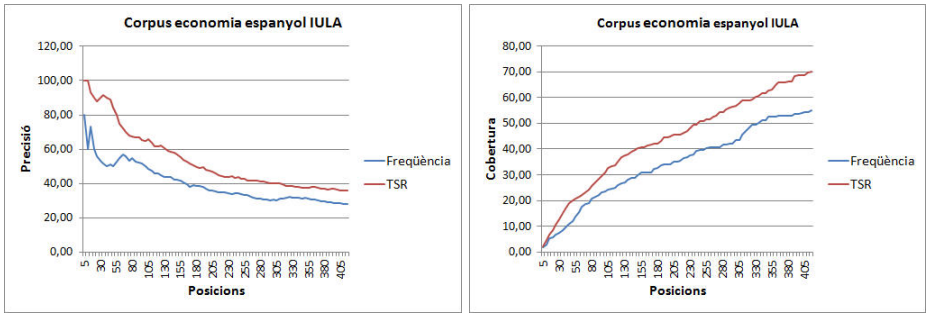


Figura 4.2: Precisió i cobertura corpus economia espanyol (IULA).

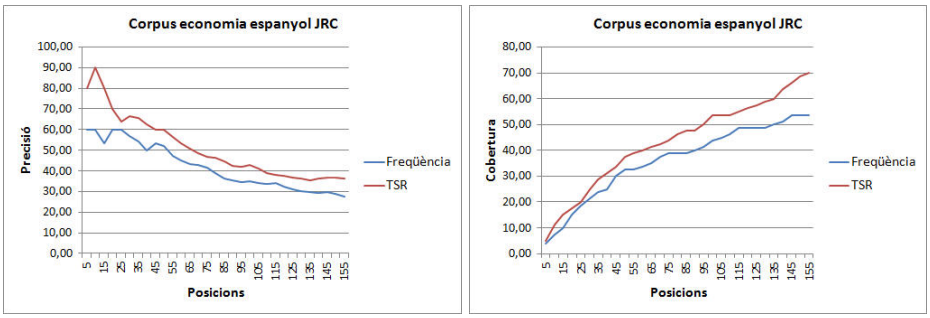


Figura 4.3: Precisió i cobertura corpus economia espanyol (JRC).

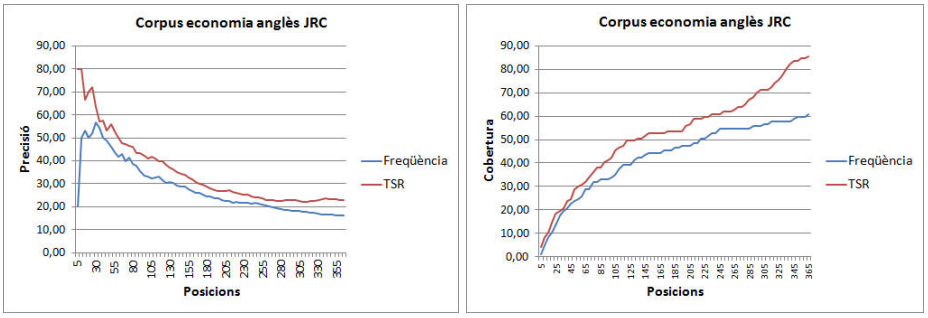


Figura 4.4: Precisió i cobertura corpus economia anglès (JRC).

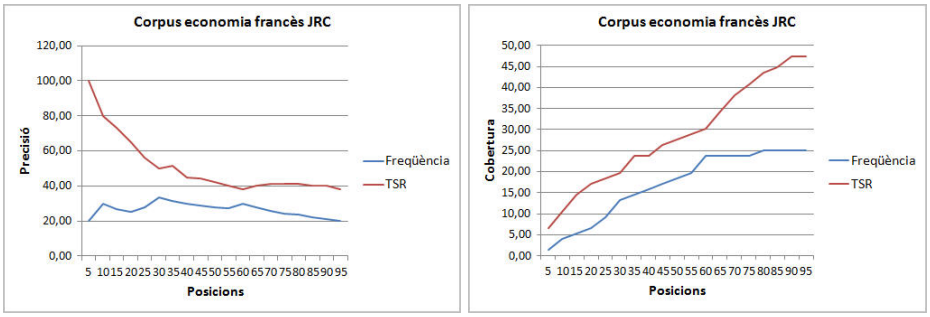


Figura 4.5: Precisió i cobertura corpus economia francès (JRC).

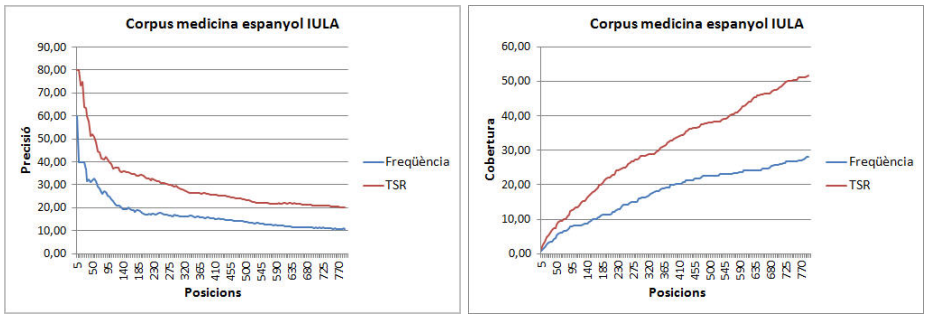


Figura 4.6: Precisió i cobertura corpus medicina espanyol (IULA).

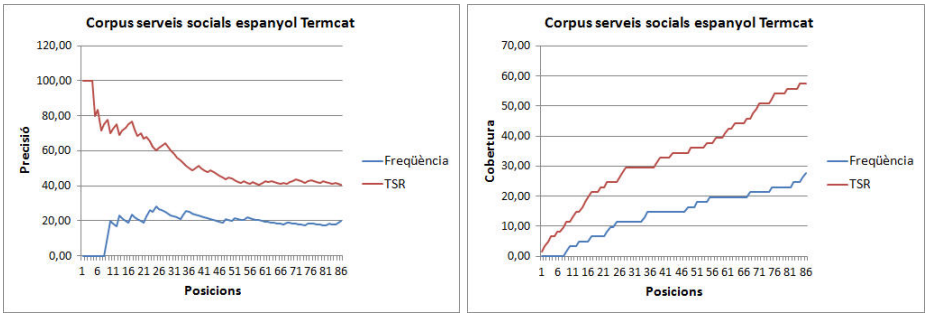


Figura 4.7: Precisió i cobertura corpus serveis socials espanyol (Termcat).

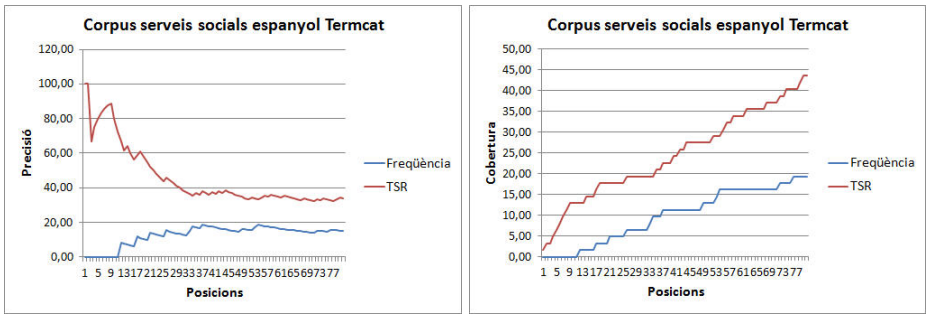


Figura 4.8: Precisió i cobertura corpus serveis socials català (Termcat).

A continuació oferim el detall de la capacitat que tenen els termes presents en el corpus d'aprenentatge d'extreure nous termes del corpus de prova. Les dades que mostrem, pertanyen als corpus d'economia, serveis socials i medicina en anglès, frànces i espanyol, i les classifiquem tenint en compte els *tokens* terminològics a partir dels quals s'han format.

Corpus d'economia en anglès del JRC
El corpus d'entrenament consta de 74 termes a partir dels quals el mètode TSR recupera 83 termes del corpus de prova d'un total de 97 termes i assoleix una cobertura del 85,57%. Els nous termes extrets són formats per *tokens* terminològics en posició inicial, final i en totes dues posicions. Així mateix, cal destacar la capacitat del mètode TSR d'identificar nous termes a mesura que s'amplia la llista de termes de referència i es duu a terme l'extracció recursiva de candidats. A continuació mostrem com queden classificats els nous termes extrets d'aquest corpus, segons la distribució de *tokens* terminològics que tenen. Com podem observar, la distribució dels termes en cada categoria és força homogènia, motiu pel qual no hi ha limitació en la posició que ocupin els *tokens* terminològics per a poder identificar termes.

a) Termes formats per una combinació de *tokens* terminològics

economic growth, economic sustainability, employment creation, employment growth, employment policies, macroeconomic stability, social partners, structural policies

b) Termes constituïts per un sol *token* terminològic situat en la posició inicial del terme de referència

budgetary savings, cyclical fluctuations, economic reform, economic shocks, economic strategy, fiscal authorities, higher productivity, industrial base, macroeconomic dialogue, price competitiveness, public services, social costs, tax systems

c) Termes constituïts per un sol *token* terminològic situat en la posició final del terme de referència

benefit reforms, competitive markets, eco-efficient technologies, effective policy, energy prices, environment-friendly technologies, global economy, inflation pressures, innovation policies, internal demand, internal market, knowledge economy, liberalised markets, stock market, trade protection

c) Termes identificats en les diferents iteracions del mètode TSR

asymmetric shocks, energy efficiency, hourly productivity, knowledge transfer, trade rules, wage-bargaining systems

Corpus d'economia en francès del JRC

El corpus d'entrenament consta de 50 termes a partir dels quals el mètode TSR recupera 36 termes del corpus de prova d'un total de 76 termes i assoleix una cobertura del 47,37%. En aquest corpus, els nous termes extrets són formats per *tokens* terminològics en posició inicial, final i un en totes dues posicions. En aquest corpus hi ha un predomini dels termes formats per *tokens* terminològics en posició final. Així mateix, cal notar que en tractar-se d'un corpus de mida més reduïda, s'observa una menor capacitat per a identificar nous termes.

a) Termes formats per una combinació de *tokens* terminològics

viabilité budgétaire

b) Termes constituïts per un sol *token* terminològic situat en la posició inicial del terme de referència

commerce international, mesures incitatives, politique industrielle

c) Termes constituïts per un sol *token* terminològic situat en la posició final del terme de referència

avance économique, coûts sociaux, coûts économiques, durabilité sociale, intégration sociale, mobilité transnationale, redressement économique, redressements économiques, secteurs économiques

Corpus de serveis socials en espanyol del Termcat

El corpus d'entrenament consta de 31 termes a partir dels quals el mètode TSR recupera 35 termes del corpus de prova d'un total de 61 termes i assoleix una cobertura del 57,38%. En aquest corpus, hi ha un predomini de termes extrets a partir de *tokens* terminològics situats en posició inicial. Amb tot, hem classificat termes formats per *tokens* terminològics en posicions inicial i final, en posició final i també com a resultat del procés recursiu d'extracció de termes.

a) Termes formats per una combinació de *tokens* terminològics

economic growth, economic sustainability, employment creation, employment growth, employment policies, macroeconomic stability, social partners, structural policies

b) Termes constituïts per un sol *token* terminològic situat en la posició inicial del terme de referència

asistencia tecnológica, atención integral, atención posadoptiva, atención precoz, atención residencial, discapacidad intelectual, discapacidad sensorial, inserción laboral, protección jurídica, violencia machista

c) Termes constituïts per un sol *token* terminològic situat en la posició final del terme de referència

asistente personal, club social, comedor social, integración familiar, prestación social, teleasistencia domiciliaria, trabajadora social

c) Termes identificats en les diferents iteracions del mètode TSR

centro residencial, integración laboral

Corpus de medicina en espanyol de l'IULA

El corpus d'entrenament consta de 181 termes a partir dels quals el mètode TSR recupera 159 termes del corpus de prova d'un total de 307 termes i assoleix una cobertura del 51,79%. Del corpus de prova s'identifiquen termes formats per *tokens* terminològics en posició inicial, final, en totes dues posicions i també termes en el procés iteratiu d'extracció. En aquest cas, s'observa un major nombre de termes formats per *tokens* terminològics en posició final.

a) Termes formats per una combinació de *tokens* terminològics

bronquitis asmatiforme, enfermedad aguda, enfermedad cardiaca, enfermedad crónica, enfermedad inflamatoria, enfisema pulmonar, estado basal, estudio funcional, evaluación ambulatoria, evaluación basal, fase estable, infección secundaria, insuficiencia renal, mucosa nasal, prueba funcional, rinitis crónica, sistema vagal, terapia broncodilatadora, terapia inhalatoria, tratamiento oral, volumen pulmonar

b) Termes constituïts per un sol *token* terminològic situat en la posició inicial del terme de referència

actividad fagocítica, asma corticorresistente, broncoespasmo intraoperatorio, bronquitis industrial, enfermedad coronaria, estudio inmunoalergológico, estudio otorrinológico, examen endoscópico, infección bacteriana, infección helmíntica, infección recurrente, infección urinaria, musculatura traqueal, obstrucción irreversible, prueba intradérmica, reactividad

sanguínea, respiración bucal, respuesta fisiológica, respuesta pupilar, rinitis activa, ritmo circadiano, sistema inmunitario, sistema respiratorio, tejido cerebral, tejido ectodérmico, tos diurna, tos emetizante

c) Termes constituïts per un sol *token* terminològic situat en la posició final del terme de referència

adrenoleucodistrofia infantil, afección respiratoria, afectación bronquial, afectación nutricional, agente etiológico, agente farmacológico, alergia rinítica, alotransplante pulmonar, aminofilina intravenosa, anestesia general, anestesia inhalatoria, antimúsculo liso, aparato respiratorio, biopsia nasal, biopsia renal, capacidad cardíaca, cinta ergométrica, circulación periférica, congestión nasal, control inmunológico, dermatitis eczematosa, eosinofilia periférica, esteroide oral, estimulación vagal, fibrosis pulmonar, fisioterapia pulmonar, hepatitis crónica, hipereosinofilia periférica, irritación nasal, microaspiración alimentaria, moco nasal, morbilidad respiratoria, origen viral, poliposis nasal, provocación nasal, prurito nasal, quejido respiratorio, salud infantil, secreción nasal, urticaria crónica

c) Termes identificats en les diferents iteracions del mètode TSR

agente colinérgico, aminofilina endovenosa, anestesia conductiva, antibioticoterapia parenteral, aparato locomotor, aparato nasopulmonar, capacidad fagocítica, eosinofilia sanguínea, esteroide parenteral, fibrosis quística, inmunoglobulina endovenosa, otitis recurrente, sepsis urinaria

La tipologia de termes que extreu el mètode TSR a partir de *tokens* terminològics de referència en posició inicial, en posició final i en totes dues posicions és idèntica en els quatre corpus analitzats, encara que es nota una major presència de termes quan aquests estan constituïts per *tokens* terminològics en posició final. Així mateix, durant el procés recursiu

d'extracció de termes són tres els corpus especialitzats en els quals s'identifiquen termes, ja que del corpus d'economia en francès no se n'identifica cap; alhora que observem una major capacitat de creació de nous termes en el corpus d'economia en anglès respecte de la resta de corpus. Així, doncs, tenint en compte que els termes identificats pertanyen a quatre corpus que tenen mides diferents i són d'àmbits d'especialitat i llengües diferents, podem constatar que el mètode TSR és adequat per a poder-se implementar sense limitació en la mida del corpus i en dominis temàtics variats.

Anàlisi qualitativa dels resultats

L'anàlisi qualitativa dels resultats obtinguts amb l'extracció de termes per mitjà del mètode TSR i el mètode freqüència té com a objectiu completar l'avaluació quantitativa de les dades que acabem de presentar, per tal de poder observar quins termes són extrets amb cada mètode i analitzar per què es produeixen diferències en els resultats d'extracció. El conjunt de termes que analitzem pertanyen als corpus d'economia en anglès i francès, al corpus de serveis socials en espanyol i al corpus de medicina en espanyol. De cada corpus recollim els termes que recupera cada un dels mètodes al final del procés recursiu d'extracció de candidats a terme, i també aportem els termes que identifica solament el mètode freqüència en el conjunt de candidats que extreu del corpus de prova d'aquests quatre dominis d'especialitat.

El conjunt de termes del corpus d'*economia en anglès* de JRC que recollim en la taula 4.19 han estat extrets dels 365 candidats obtinguts en el procés d'extracció recursiva de termes. Constatem que en aquesta llista de candidats hi ha un alt predomini de termes extrets amb el mètode TSR, comparant-ho amb el mateix nombre de candidats extrets amb el mètode freqüència. Els 36 termes recuperats solament amb el mètode TSR estan formats per *tokens* terminològics procedents de la llista de termes de referència del corpus d'entrenament; en canvi, els 4 termes extrets amb el mètode freqüència no contenen cap d'aquests *tokens* terminològics.

La llista de termes que solament identifica el mètode TSR mostra que l'extracció per mitjà de *tokens* terminològics permet concentrar un alt nombre de termes en cada iteració fins al final del procés recursiu; a diferència del que succeeix amb el mètode freqüència, que distribueix els termes en una llista llarga de candidats segons la freqüència d'aparició que tingui la unitat terminòlogica en el corpus especialitzat, motiu pel qual extreu un menor nombre de termes en cada iteració i al final del procés recursiu.

Amb el detall de termes recuperats per cada mètode, hem comprovat si els termes presents solament a freqüència també poden ser identificats amb el mètode TSR revisant manualment la llista de 365 candidats extrets en el procés iteratiu. En aquest sentit, hem constatat que dels 4 termes extrets amb el mètode freqüència el terme *market-based instruments* sí que es pot recuperar i que els termes *climate change*, *entrepreneurial culture* i *local governments* no es poden recuperar, per no estar formats per *tokens* terminològics presents ni en els termes de referència ni en els nous termes extrets dels candidats.

A més d'analitzar els termes extrets únicament pel mètode freqüència al final del procés iteratiu, hem volgut anar al detall dels termes que recupera aquest mètode en el conjunt de candidats que extreu del corpus de prova i comparar-los amb els termes que extreu el mètode TSR. Concretament, el mètode freqüència extreu un total de 1.591 candidats del corpus de prova del domini d'economia en anglès i recupera un total de 13 termes, incloent-hi els 4 termes abans esmentats, que no són extrets pel mètode TSR en el procés iteratiu. Hem comprovat si aquests termes poden ser extrets pel mètode TSR fent la revisió manual dels 365 candidats a terme, i els resultats que hem obtingut són els següents:

- Termes que pot extreure el mètode TSR

  business investment, civil society, common goods, geographical mobility, innovative activity, market-based instruments, national regulation, rates price, researcher mobility, rural development

- Termes que no pot extreure el mètode TSR

  climate change, entrepreneurial culture, local governments

Observem que revisant manualment la llista de 365 candidats a terme, el mètode TSR pot recuperar 10 del total de 13 termes que identifica el mètode freqüència en el conjunt de 1.591 candidats a termes i que no són recollits inicialment en el procés iteratiu. Així, doncs, podem confirmar que en el corpus d'economia en anglès de JRC pràcticament la totalitat de termes extrets amb el mètode freqüència també poden ser recuperats pel mètode TSR. És rellevant destacar en aquest cas que els resultats amb el mètode TSR s'obtenen a partir d'una llista de termes de referència presents en el corpus d'entrenament i es presenten en una llista signitivament més reduïda de candidats a terme que la del mètode freqüència.

Taula 4.19: Termes identificats amb TSR i freqüència (I).

Economia anglès JRC	
Mètode TSR	
asymmetric shocks	innovation policies
benefit reforms	internal demand
budgetary savings	knowledge economy
competitive economy	knowledge transfer
cyclical fluctuations	labour supply
eco-efficient technologies	liberalised markets
economic conditions	lifecycle approach
economic partners	long-term sustainability
economic reform	macroeconomic dialogue
economic slowdown	microeconomic policies
economic sustainability	price competitiveness
effective policy	public services
environment-friendly technologies	social costs
environmental damage	social inclusion
environmental sustainability	stock market
female employment	trade protection
fiscal authorities	trade rules
inflation pressures	wage-bargaining systems
Mètode freqüència	
climate change	
entrepreneurial culture	
local governments	
market-based instruments	

Si analitzem el corpus *economia en francès* de JRC, observem que els termes recollits en la taula 4.20 han estat extrets dels 96 candidats obtinguts al final del procés recursiu d'extracció de termes. En aquest corpus el mètode TSR aconsegueix identificar 20 termes i el mètode freqüència

n'identifica 4. En aquest corpus, els termes identificats amb el mètode freqüència tampoc no estan formats per *tokens* terminològics presents en el corpus d'entrenament. Hem comprovat si revisant manualment els 96 candidats extrets al final del procés recursiu, els 4 termes extrets amb el mètode freqüència també poden ser identificats pel mètode TSR. En aquest cas, hem constatat que amb aquesta revisió manual el mètode TSR no pot extreu cap dels termes que recull freqüència.

Així mateix, hem recollit els termes que recupera solament el mètode freqüència en el conjunt de candidats que extreu del corpus de prova. El mètode freqüència extreu un total de 820 candidats del corpus de prova, en els quals hi ha 18 termes, incloent-hi els 4 termes que recupera en el procés recursiu, que no extreu el mètode TSR. En aquest sentit, hem comprovat si revisant manualment els 96 candidats resultants del procés recursiu aquests 18 termes poden ser recuperats pel mètode TSR. Els resultats obtinguts són els següents:

- Termes que pot extreure el mètode TSR

  excellence européenne, investissement public

- Termes que no pot extreure el mètode TSR

  avantages compétitifs, efficacité énergétique, financement national, fonds structurels, guichets uniques, innovation technologique, investissement public, investissements étrangers, jeunes entrepreneurs, marché intérieur, marché unique, plan extérieur, plans stratégiques, productivité horaire, recherche publique, services publics, transfert transfrontalier

Constatem que dels 18 termes que extreu freqüència, únicament *excellence européenne* i *investissement public* poden ser recuperats pel mètode TSR i, en conseqüència, una menor capacitat per identificar la totalitat de termes extrets amb el mètode freqüència. Amb tot, si comparem el corpus d'economia en francès amb el d'anglès, observem que la mida del primer corpus és el doble que la del segon i, alhora, que el nombre de termes

de referència presents en el corpus d'entrenament és inferior. Aquesta combinació de factors pot influir en una menor capacitat del mètode TSR d'extreure aquests 18 termes.

Taula 4.20: Termes identificats amb TSR i freqüència (II).

Economia francès JRC	
Mètode TSR	
activité économique	mobilité professionnelle
avance économique	mobilité transnationale
chocs économiques	politique budgétaire
commerce international	politique industrielle
coûts sociaux	politique économique
coûts économiques	redressement économique
durabilité sociale	stabilité macroéconomique
finances publiques	stabilité économique
intégration sociale	viabilité budgétaire
mesures incitatives	viabilité financière
Mètode freqüència	
avantages compétitifs	
efficacité énergétique	
marché intérieur	
productivité horaire	

Amb relació al corpus de *serveis socials en espanyol* del Termcat, en la taula 4.21 recollim els termes extrets dels 86 candidats obtinguts al final del procés d'extracció recursiva de termes. El mètode TSR recupera 26 termes, els quals estan formats pels *tokens* terminològics dels termes de referència del corpus d'entrenament, per davant dels 6 termes recuperats per freqüència, que no estan formats per cap dels *tokens* terminològics.

Si observem la capacitat que té el mètode TSR per recuperar els 6 termes presents a freqüència revisant manualment els 86 candidats, constatem que no en pot recuperar cap.

Així mateix, analitzem quins termes extreu únicament el mètode freqüència en el conjunt de candidats que recupera del corpus de prova. El mètode freqüència extreu un total de 911 candidats a terme del corpus de prova d'aquest domini temàtic, entre els quals identifiquem 7 termes que no extreu el mètode TSR durant el procés recursiu. Per aquest motiu, hem comprovat si revisant manualment els 86 candidats a terme obtinguts en aquest procés recursiu, el mètode TSR pot extreure aquests 7 termes. Hem obtingut els resultats següents:

- Termes que pot extreure el mètode TSR

  cap

- Termes que no pot extreure el mètode TSR

  autorización administrativa, enfermedad mental, equilibrio territorial, hogar residencia, personas mayores, piso puente, terapia ocupacional

En aquest corpus, el mètode TSR no pot extreure cap dels 7 temres que recupera el mètode freqüència. Els factors que poden haver influït en la capacitat del mètode TSR a no poder identificar aquests termes són la mida del corpus de prova, que és gairebé el doble de gran que el del corpus d'economia en francès, i el nombre de termes de referència presents en el corpus d'entrenament, que és inferior al del corpus d'economia en francès.

Taula 4.21: Termes identificats amb TSR i freqüència (III).

Serveis socials espanyol Termcat	
Mètode TSR	
acceso universal	inserción laboral
asistencia tecnológica	integración familiar
atención diurna	integración laboral
atención integral	participación ciudadana
atención posadoptiva	protección jurídica
atención precoz	responsabilidad pública
atención residencial	riesgo social
atención social	sensibilización social
centro residencial	teleasistencia domiciliaria
club social	titularidad pública
comedor social	trabajadora social
discapacidad sensorial	trabajo social
exclusión social	urgencia social
Mètode freqüència	
autorización administrativa	
enfermedad mental	
hogar residencia	
personas mayores	
plan sectorial	
terapia ocupacional	

El conjunt de termes que recollim en les taules 4.22 i 4.23 pertanyen al corpus de prova del domini temàtic de *medicina en espanyol* de l'IULA. En el procés recursiu d'extracció de termes del mètode TSR s'obtenen 587 candidats, dels quals 90 són termes extrets sols amb el mètode TSR i formats per *tokens* terminològics dels termes del corpus d'entrenament, i 18 són identificats pel mètode freqüència els *tokens* terminològics dels quals no són coincidents amb els dels termes extrets pel mètode TSR.

Com hem fet en els corpus precedents, comprovem si aquests 18 termes extrets per freqüència poden ser recuperats amb el mètode TSR revisant manualment els 587 candidats extrets en el procés recursiu. Hem constatat que els termes *antígeno alimentario, componente inflamatorio, conducta terapéutica, corte transversal, intervención quirúrgica, modalidad terapéutica, período intercrítico, período perinatal* també poden ser recuperats pel mètode TSR i que els termes *acción antiinflamatoria, agonista beta, apnea voluntaria, decúbito supino, hipersecreción mucosa, nedocromil sódico, proteína catiónica, pulmón contralateral, tabaquismo pasivo, vacunación antigripal* no es podrien recuperar per estar formats per *tokens* terminològics no coincidents ni amb els que formen part dels termes de referència del corpus d'entrenament ni tampoc dels nous termes extrets.

Completem l'anàlisi amb els termes que el mètode freqüència indentifica en el conjunt de candidats que extreu del corpus de prova. El mètode freqüència extreu d'aquest domini d'especialitat un total de 4.896 candidats a terme, en els quals hi ha 66 termes extrets únicament per freqüència. Hem comprovat si revisant manualment els 587 candidats extrets en el procés recursiu, el mètode TSR també pot extreure aquests termes. Hem obtingut els resultats següents, en els quals també incloem els termes identificats per freqüència en el procés recursiu.

- Termes que pot extreure el mètode TSR

  antígeno alimentario, albúmina humana, componente inflamatorio, conducta terapéutica, contraindicación quirúrgica, corte transversal, cultivo negativo, depresión nerviosa, edad gestacional, enuresis nocturna, estridor inspiratorio, infiltración celular, intervención quirúrgica, modalidad terapéutica, mortalidad fetal, patogenia inmunológica, penicilina sódica, perfil humoral, período intercrítico, período perinatal, régimen terapéutico, sibilancia nocturna

- Termes que no pot extreure el mètode TSR

  acción antiinflamatoria, actividad broncodilatadora, agonista beta, alta hospitalaria, apnea voluntaria, artritis reumatoide, barrera fetoplacentaria, barrera gastrointestinal, cadena lipofílica, candidiasis orofaríngea, carga hereditaria, colinomimético indirecto, comunicación interventricular, corticoterapia sistémica, decúbito dorsal, decúbito supino, esqueleto facial, filtración glomerular, forma lobulillar, glutamato monosódico, hipersecreción mucosa, madre adolescente, malformación congénita, mixoma embolizante, nedocromil sódico, paro cardiorrespiratorio, perforación distal, peso fetal, polinosis dermatológica, proteinosis alveolar, proteína catiónica, pulmón contralateral, pulmón derecho, punción venosa, radioterapia superficial, retraso mental, ruptura uterina, tabaquismo pasivo, talla fetal, tumor maligno, vacunación antigripal, ventilador volumétrico, verruga plana

Es constata que d'un total de 66 termes extrets per freqüència en el conjunt total de 4.896 candidats extrets del corpus de prova del domini de medicina, el mètode TSR pot extreure 22 termes revisant manualment els 587 candidats extrets en el procés recursiu. Una mida gran del corpus, juntament amb una llista llarga de termes de referència en el corpus d'entrenament, han permès poder identificar una part dels termes presents únicament a freqüència. La resta de termes no es poden recuperar perquè estan formats per *tokens* terminològics no inclosos ni en la llista de termes de referència ni en els termes nous.

Taula 4.22: Termes identificats amb TSR i freqüència (IVa).

Medicina espanyol IULA	
Mètode TSR	
actividad broncodilatadora	enfermedad cardiaca
adrenoleucodistrofia infantil	eosinofilia periférica
afección respiratoria	eosinofilia sanguínea
afectación bronquial	espasmo bronquial
afectación nutricional	estado basal
agente colinérgico	estado nutricional
agente etiológico	esteroide oral
agente farmacológico	esteroide parenteral
alergia rinítica	estimulación vagal
alotransplante pulmonar	estudio funcional
anemia falciforme	estudio inmunoalergológico
anestesia inhalatoria	estudio otorrinológico
antibioticoterapia parenteral	examen endoscópico
antimúsculo liso	fibrosis pulmonar
aparato locomotor	fibrosis quística
aparato nasopulmonar	fisioterapia pulmonar
aparato respiratorio	hepatitis crónica
asma episódica	hipereosinofilia periférica
asma grave	hiperirritabilidad bronquial
asma intrínseca	infección bacteriana
biopsia nasal	infección secundaria
biopsia renal	infección urinaria
broncoespasmo intraoperatorio	inmunoglobulina endovenosa
bronquitis industrial	insuficiencia cardiaca
capacidad cardíaca	irritación nasal
capacidad fagocítica	mecanismo inmunológico
cinta ergométrica	microaspiración alimentaria
circulación periférica	mucosa nasal
congestión nasal	musculatura traqueal
control inmunológico	obstrucción irreversible
dermatitis eczematosa	origen viral
dificultad respiratoria	otitis recurrente
dolor abdominal	prueba funcional

Taula 4.23: Termes identificats amb TSR i freqüència (IVb).

Medicina espanyol IULA	
Mètode TSR	
prueba intradérmica	acción antiinflamatoria
prurito nasal	agonista beta
quejido respiratorio	antígeno alimentario
reactividad sanguínea	apnea voluntaria
rinitis activa	componente inflamatorio
rinitis crónica	conducta terapéutica
ritmo circadiano	corte transversal
salud infantil	decúbito supino
sensibilidad diagnóstica	hipersecreción mucosa
sepsis urinaria	intervención quirúrgica
sistema vagal	modalidad terapéutica
sobredistensión pulmonar	nedocromil sódico
tejido cerebral	período intercrítico
tejido ectodérmico	período perinatal
tensión arterial	proteína catiónica
terapia broncodilatadora	pulmón contralateral
terapia inhalatoria	tabaquismo pasivo
tos crónica	vacunación antigripal
tos diurna	
tos emetizante	
tratamiento oral	
urticaria crónica	
valor basal	
Mètode freqüència	
acción antiinflamatoria	intervención quirúrgica
agonista beta	modalidad terapéutica
antígeno alimentario	nedocromil sódico
apnea voluntaria	período intercrítico
componente inflamatorio	período perinatal
conducta terapéutica	proteína catiónica
corte transversal	pulmón contralateral
decúbito supino	tabaquismo pasivo
hipersecreción mucosa	vacunación antigripal

Després de fer l'anàlisi qualitativa dels termes corresponents a quatre dominis d'especialitat, en tres llengües i amb volum de corpus diferent, podem constatar amb matisacions la hipòtesi inicial que plantejàvem en referència al fet que un mètode recursiu d'extracció automàtica de termes basat en estratègies estadístiques, i concretament amb *tokens* terminològics, permet extreure un major nombre de termes que un mètode no recursiu basat en la freqüència. Cal matisar aquesta hipòtesi inicial dient que es recuperen més termes tenint en compte el nombre de candidats extrets per ambdós mètodes, és a dir, el mètode recursiu aconsegueix concentrar en un nombre reduït de candidats una quantitat molt més elevada de termes que no pas el mètode no recursiu basat en la freqüència. És així perquè l'extracció de termes basada en freqüència distribueix els termes entre tots els candidats extrets, considerant únicament el nivell d'aparició que els termes tinguin en el corpus com a paràmetre per a endreçar-los en la llista de resultats, i en conseqüència té menor capacitat d'identificació de termes en un segment reduït de candidats. Aquesta hipòtesi es confirma en tots els corpus analitzats.

Així mateix, hem completat l'anàlisi qualitativa de les dades identificant els termes presents en els candidats dels corpus de prova extrets pel mètode de freqüència i que no són inicialment presents en els resultats del mètode TSR. Tenint en compte que el mètode TSR extreu els termes a partir dels *tokens* terminològics dels corpus d'entrenament i dels nous termes extrets, i que el mètode freqüència utilitza tots els candidats dels corpus de prova, hem volgut constatar la capacitat d'extracció de termes del mètode TSR. En aquest sentit, el mètode TSR identifica 10 dels 13 termes que el mètode freqüència extreu del corpus de prova d'economia en anglès, recupera 2 dels 18 termes del corpus d'economia en francès, no pot identificar cap dels 7 termes que extreu freqüència del corpus de serveis socials en espanyol i extreu 22 dels 66 termes de freqúència presents en el corpus de medicina en espanyal. Tenint en compte el nivell de detall de l'anàlisi, els resultats mostren una bona capacitat d'extracció de termes del mètode TSR. La major o menor capacitat del mètode TSR en cada corpus d'incorporar els termes presents únicament als resultats de

freqüència té una relació directa amb la mida del corpus de prova i amb el nombre de termes de referència presents en el corpus d'entrenament. És una limitació que convé tenir en compte a l'hora d'implementar aquest mètode en altres corpus. Amb tot, considerem que el mètode TSR pot arribar a recuperar tots els termes de freqüència fent l'extracció recursiva dels candidats dels corpus de prova amb els termes de referència presents en aquests corpus.

Observant la capacitat d'extracció de termes per corpus que mostra el mètode TSR, podem afirmar que com més reduïda és la llista de termes de referència menys capacitat té d'extreure termes, com passa en els corpus d'economia en francès i de serveis socials. En el corpus d'economia en anglès pràcticament recupera tots els termes de freqüència. I en el corpus de medicina constatem que, malgrat disposar d'una àmplia llista de termes de referència, el mètode TSR extreu un nombre significatiu de termes, però no arriba a fer-ho en la seva totalitat. Aquest nivell de detecció del mètode TSR pel que fa als termes del mètode freqüència està influït pel fet que l'extracció es fa a partir dels termes de referència presents en el corpus d'entrenament; en canvi, el mètode freqüència en els casos que hem mostrat, pot extreure aquests termes perquè no està subjecte a cap limitació a l'hora d'extreure els candidats a terme dels corpus de prova.

Dels exemples recollits en l'anàlisi qualitativa es desprèn l'avantatge que representa la implementació del mètode TSR respecte del mètode de freqüència amb relació al nombre de candidats que extreu del corpus i el conjunt de termes que en recupera. Aquesta concentració dels termes en un nombre reduït de candidats facilita la tasca de revisió manual de les unitats extretes en temps i en redueix la inversió econòmica. La distribució dels termes en l'extens conjunt de candidats que s'extreu amb el mètode freqüència dificulta molt aquesta tasca i encareix el procés d'obtenció de termes aplicant estratègies automàtiques.

Els termes presents en els corpus de prova en cada un dels dominis temàtics analitzats, són identificats pel mètode TSR i el mètode freqüència

amb estratègies diferents: d'una banda, el mètode TSR en fa l'extracció per mitjà de *tokens* terminològics, que permet recollir pràcticament la totalitat de termes del corpus de prova en una llista reduïda de candidats a termes; d'altra banda, el mètode freqüència també identifica els termes d'aquests corpus, però en tenir en compte solament la freqüència d'aparició dels candidats en el corpus, la llista de resultats és extensa i demana una revisió llarga i costosa dels candidats proposats. Partint d'aquesta premissa inicial, amb el mètode TSR es dóna prioritat als termes que estan formats parcialment o totalment per unitats terminològiques; en canvi, en el mètode freqüència els candidats tenen major o menor pes segons el nombre de vegades que apareixen en el corpus.

La contribució que fa el mètode TSR a la tasca d'extracció automàtica de termes en corpus especialitzat és la selecció de candidats per mitjà de *tokens* terminològics per tal de classificar els resultats obtinguts segons la proximitat que tenen amb els termes propis de l'àmbit del corpus i desestimar les unitats que no tenen un component terminològic. A més, aquest model de processament permet focalitzar l'avaluació dels resultats finals en un grup reduït de candidats. Així mateix, aquest mètode és generalitzable a altres àmbits d'especialitat i permet processar corpus de menor o major volum, tal com indiquen els resultats obtinguts. I té capacitat per a processar corpus de llengües que tenen una tipologia flexional, perquè no s'han d'adaptar patrons lingüístics en el procés d'extracció terminològica.

En definitiva, constatem que l'extracció recursiva de *tokens* terminològics del mètode TSR permet recuperar els termes presents en un domini d'especialitat *a)* concentrant-los en un nombre de candidats reduït i *b)* identificant bona part dels termes que extreu el mètode freqüència. El filtratge per iteracions té l'avanatatge de concentrar en un conjunt redüit de candidats la major part de termes d'un corpus. Per contra, el mètode freqüència *a)* identifica un nombre baix de termes en cada una de les iteracions del procés recursiu i *b)* extreu una llarga llista de candidats del corpus, fet que dificulta la localització dels termes que conté.

4.5 Conclusions

En el present capítol hem descrit una proposta experimental d'extracció automàtica de termes presents en corpus especialitzats basada en un mètode no supervisat que permet l'extracció recursiva de candidats a terme per mitjà de *tokens* terminològics i que anomenem mètode *token slot recognition* (mètode TSR). En aquest sentit, hem elaborat un algorisme d'extracció de candidats basat en una estratègia estadística, amb l'objectiu de poder ser implementat en corpus especialitzats de diferent volum, que pertanyin a diferents àmbits d'especialitat i que estiguin disponibles en diverses llengües.

El procés d'extracció de candidats a terme del mètode TSR se centra en quatre passos: 1) selecció automàtica dels candidats presents en un corpus especialitzat; 2) identificació dels termes que són presents en el corpus, ja sigui de forma manual o a partir d'una llista prèvia de termes de referencia, que servirà de base per al mètode TSR; 3) filtratge recursiu dels candidats a terme formats per *tokens* terminològics; 4) selecció manual final dels candidats que són termes.

El mètode TSR ha estat provat en corpus de l'àmbit de l'economia, els serveis socials i la medicina en anglès, francès, espanyol i català, els quals tenen un volum en nombre de paraules força diferent. L'ús d'aquesta varietat de corpus ha permès extreure conclusions en detall del rendiment del mètode TSR que hem presentat.

Els resultats obtinguts amb el mètode TSR en set corpus especialitzats han estat contrastats amb els que s'obtenen aplicant-hi el mètode de freqüència. Ambdós mètodes han estat comparats segons els resultats que s'obtenen en diferents posicions (seleccionades aleatòriament) en les quals queden situats els candidats a terme i els resultats que s'obtenen en cada una de les iteracions en què el mètode TSR extreu candidats i identifica termes. Els resultats indiquen una major capacitat d'extracció de termes per part del mètode TSR.

L'avaluació dels resultats obtinguts amb el mètode TSR i el mètode de freqüència s'ha dut a terme a partir de les mètriques de precisió i cobertura, juntament amb la mesura-F. Per a fer-ho, els corpus especialitzats han estat dividits en corpus de prova i d'avaluació. Els candidats a terme han estat extrets del corpus d'avaluació i s'han utilitzat els termes presents en els corpus de prova per a filtrar els resultats. En els resultats de l'avaluació s'obté un percentatge de precisió i cobertura més alts amb el mètode TSR que no pas amb el de freqüència per a tots els corpus.

Els resultats obtinguts de l'avaluació d'ambdós mètodes mostren la capacitat que té el mètode TSR d'extreure un major nombre de termes que no pas el mètode de freqüència en una llista de candidats més reduïda. En aquest sentit, una de les millores que introdueix el mètode TSR respecte de les estratègies descrites en el capítol 3 és la de concentrar els termes presents en els corpus especialitzats en una llista reduïda de candidats. Així és més àgil la tasca de revisió manual final dels resultats per part d'un especialista. Amb l'objectiu de reduir al màxim el nombre de candidats que han de ser revisats manualment, en el capítol 6 proposem la combinació del mètode TSR amb mesures d'associació lèxica per tal d'assolir una major rendibilitat en l'extracció automàtica de termes.

Així mateix, el mètode TSR té capacitat per a processar corpus en diverses llengües que siguin de major o menor volum i que pertanyin a diferents àmbits d'especialitat, tal com mostren els resultats obtinguts, flexibilitat que no permeten tots els mètodes d'extracció automàtica de termes.

Finalment, convé tenir present que el mètode TSR permet de ser combinat amb estratègies lingüístiques per a poder afinar els resultats obtinguts. Així, en ser combinat, passa de ser un mètode de caràcter estadístic a convertir-se en una estratègia híbrida d'extracció de termes. Aquesta capacitat del mètode fa possible millorar el rendiment d'eines dissenyades per a l'extracció de termes que siguin de codi lliure i que, en conseqüència, permeten una actualització contínua.

Capítol 5

MESURES D'ASSOCIACIÓ LÈXICA

En el present capítol descrivim les mesures d'associació lèxica que hem implementat en la proposta experimental que presentem en el capítol 6 amb l'objectiu de poder millorar el procés de validació final de candidats a terme extrets d'un corpus d'especialitat. La selecció de mesures que incorporem en el nostre plantejament experimental es basa essencialment en la capacitat que tenen per a identificar candidats a terme bigrams, en la possibilitat que tenen de ser aplicades en la tasca d'extracció automàtica de termes i en la rendibilitat que han obtingut en estudis previs.

Les mesures d'associació lèxica les descrivim tenint en compte la relació que s'estableix entre els elements que formen part d'un candidat a terme i el nivell de ponderació que atorguen a un conjunt de candidats. En aquest sentit, hem classificat les mesures en quatre grans grups: mesures d'importància d'associació (apartat 5.2), mesures de força d'associació (apartat 5.3), mesures provinents de la teoria de la informació (apartat 5.4) i mesures heurístiques (apartat 5.5).

5.1 Introducció

Una mesura d'associació lèxica és una fórmula matemàtica que determina la força d'associació entre els *tokens*[1] que constitueixen un n-gram[2] tenint en compte la freqüència en què ocorren i coocorren aquests *tokens* en un corpus. Concretament, una mesura d'associació lèxica calcula la puntuació d'associació (*association score*) que hi ha entre *tokens* extrets d'un corpus. La puntuació d'associació que calculen les mesures pot ser considerada de tres maneres: per a estimar la importància d'associació entre els *tokens* d'un n-gram, per a obtenir el rànquing que ocupen els n-grams en un conjunt de dades i per a establir el rànquing dels n-grams que estiguin formats per un determinat *token*. Quant a la primera opció, la puntuació que calculen les mesures indica el tipus d'associació que tenen els diferents n-grams. Així, com més cohesionats estiguin els n-grams (com més alta sigui la puntuació d'associació) més probabilitat tindran de ser una unitat terminològica o una col·locació. En aquest sentit, es pot considerar que les mesures d'associació són bons heurístics per a determinar associacions de paraules rellevants. Pel que fa a la segona opció, la puntuació que calculen les mesures d'associació serveix per a endreçar els n-grams en un conjunt de resultats extrets d'un corpus d'acord amb un criteri de distinció entre *candidat acceptat* i *candidat no acceptat* a partir d'un determinat llindar. I en relació a la tercera opció d'interpretar la puntuació de les mesures, la puntuació d'associació serveix per a endreçar els n-grams a partir dels *tokens* que els componen (Evert, 2005; Pecina i Schlesinger, 2006).

Els *tokens* extrets d'un corpus queden classificats en una taula de contingència 2x2, tenint en compte la freqüència d'aparició que tenen en cada una de les posicions que ocupen. La taula 5.1 mostra la taula de contingència corresponent a bigrams i la notació estàndard que s'usa.

[1] Vegeu pàgina 64.
[2] Vegeu pàgina 67.

Taula 5.1: Taula de contingència 2x2.

	token2	$\sim$ *token2*	Total
token1	n_{11}	n_{12}	n_{1p}
$\sim$ *token1*	n_{21}	n_{22}	n_{2p}
Total	n_{p1}	n_{p2}	n_{pp}

La cel·la n_{11} correspon a la freqüència en què *token1* i *token2* apareixen junts. La cel·la n_{12} correspon a la freqüència en què *token1* es troba en primera posició i *token2* no es troba en segona posició. La cel·la n_{21} correspon a la freqüència en què *token1* no es troba en primera posició i *token2* es troba en segona posició. La cel·la n_{22} correspon a la freqüència en què ni *token1* ni *token2* apareixen en llurs respectives posicions. Les cel·les n_{1p}, n_{p1}, n_{2p} i n_{p2} corresponen als totals marginals, és a dir, al nombre de vegades que un *token* ocorre o no en primera posició o en segona posició d'un bigram. I la cel·la n_{pp} correspon al nombre total de bigrams localitzats en un corpus.

Les taules de contingència poden ser creades per a n-grams de mida *n*, tenint en compte que com més alt és el nombre de grams més complexitat tindrà la taula, perquè *n* s'incrementa i el recompte dels totals marginals s'incrementa en 2^n (McInnes, 2004; Evert, 2005).

Les dades de les taules de contingència poden ser usades per a avaluar n-grams fent servir mesures estadístiques, perquè aquestes mesures tenen en compte quins valors es poden esperar en la taula de contingència respecte dels valors que s'han observat en el corpus. Concretament, la freqüència observada d'un parell de *tokens* (*token1*, *token2*) es representa amb un valor numèric, O_{ij}, en el qual *i* i *j* representen la presència (valor=1) o absència (valor=0) de cada *token* en el bigram (taula 5.2).

Taula 5.2: Freqüències observades i marginals.

	token2	$\sim$ *token2*	Total
token1	O_{11}	O_{12}	= R1
$\sim$ *token1*	O_{21}	O_{22}	= R2
Total	= C_1	= C_2	= N

El valor O_{11} correspon al nombre de vegades que *token1* i *token2* apareixen junts, el valor O_{12} indica quantes vegades *token1* apareix amb una paraula diferent de *token2*. Les freqüències marginals corresponen a la suma de cada línia i cada columna, és a dir, O_{1p} és la suma de O_{11} i O_{12}, i així successivament. Les freqüències marginals també són anomenades *R1* i *R2* (valors totals de la fila), i *C1* i *C2* (valors totals de la columna). La suma de les freqüències marginals és O_{pp} -també pot correspondre a la notació *N*-, que és la suma del nombre total de bigrams.

Les freqüències observades per cada bigram tenen una taula paral·lela de freqüències estimades, les quals proporcionen les freqüències esperades, donada la hipòtesi nul·la que no hi ha associació entre les paraules dels bigrams, tal com queda reflectit en la taula 5.3.

Taula 5.3: Freqüències esperades i marginals.

	token2	$\sim$ *token2*	Total
token1	$E_{11} = \frac{R1C1}{N}$	$E_{12} = \frac{R1C2}{N}$	= R1
$\sim$ *token1*	$E_{21} = \frac{R2C1}{N}$	$E_{22} = \frac{R2C2}{N}$	= R2
Total	= C_1	= C_2	= N

Si no hi ha associació entre *token1* i *token2* s'espera que la probabilitat que apareguin junts sigui proporcional a la freqüència de cada un dels

tokens del bigram individualment. D'aquesta manera, si tots dos *tokens* tenen una freqüència alta, llavors es pot esperar una freqüència alta per a llur valor estimat E_{11}, i a la inversa.

Les mesures d'associació lèxica són emprades en la tasca d'extracció automàtica de terminologia (Daille, 1997; Pazienza *et al.*, 2005; Boulaknadel *et al.*, 2008) i també en la identificació de col·locacions (Krenn i Evert, 2001; Evert i Krenn, 2005), amb l'objectiu d'aprofitar la capacitat que tenen de mesurar la força d'associació que hi ha entre els *tokens* d'un n-gram.

En aquest sentit, un dels treballs que ha avaluat la rendibilitat que ofereixen les mesures en la detecció automàtica de termes és el de Schmidt (2001), en el qual fa una completa descripció de les mesures khi quadrat de Pearson, *t* de Student, informació mútua, ràtio log-likelihood, C-Value aplicades a aquesta tasca. Els resultats obtinguts indiquen que informació mútua i prova X^2 són útils per a corpus grans.

Així mateix, en Pazienza *et al.* (2005) hi ha un ampli estudi comparatiu de mesures en el qual intervenen la freqüència, *t* de Student, informació mútua, MI^3, coeficient Dice, ràtio log-likelihood, C-Value i Co-occurrence, i són les mesures *t* de Student i log-likelihood les que obtenen més bons resultats.

En relació als estudis que s'han dut a terme per a avaluar la rendibilitat l'ús de les mesures d'associació lèxica en la detecció de col·locacions, convé destacar l'anàlisi comparativa que va fer Thanopoulos *et al.* (2002) de les mesures *t* de Student, prova khi quadrat de Pearson, ràtio log-likelihood, pointwise mutual information, freqüència i una mesura provinent de la teoria de la informació, *mutual dependency*. Aquesta anàlisi té com a objectiu identificar associacions lèxiques formades per bigrams fent servir dos estàndards de referència, wordnet i llistes d'entitats amb nom (*name entity*). En l'avaluació comparativa a partir d'aquests dos estàndards, la ràtio log-likelihood, *mutual dependency* i la prova X^2 de

Pearson obtenen més bons resultats que no pas *t* de Student, freqüència i pointwise mutual information.

En l'estudi d'Evert i Krenn (2005) es duu a terme una avaluació empírica de les mesures d'associació estadística *t* de Student, ràtio log-likelihood, freqüència i la prova khi quadrat de Pearson (amb l'aplicació de la correcció de Yates) per a l'extracció de col·locacions d'un corpus. Els resultats obtinguts d'aquest estudi indiquen que *t* de Student i ràtio log-likelihood ofereixen més bons resultats que no pas la prova X^2 de Pearson i la freqüència.

En l'article de Pecina (2010) s'avaluen vuitanta-dues mesures estadístiques aplicades a l'extracció de col·locacions bigrams. Els resultats obtinguts s'avaluen calculant la precisió i la cobertura i també la precisió mitjana (*mean average precision*). L'estudi conclou que no és possible seleccionar una única mesura universal per a la identificació de col·locacions i que les mesures ofereixen diferents resultats per a diferents tasques, segons les dades de partida, les llengües de treball i el tipus de col·locació en què se centri la tasca que es vol fer. Les mesures que implementem en la nostra proposta experimental i que obtenen millors resultats en aquest estudi són pointwise mutual information, prova khi quadrat de Pearson, ràtio odds i coeficient Jaccard.

I en el treball de Lyse i Andersen (2012) s'apliquen nou mesures estadístiques per a l'extracció de bigrams (coeficient Dice, coeficient Jaccard, *t* de Student, ràtio log-likelihood, pointwise mutual information, Z-score, Z-score corregit, prova khi quadrat de Pearson, ràtio odds) i quatre mesures estadístiques per a l'extracció de trigrams (informació mútua, ràtio log-likelihood, pointwise mutual information, Poisson-Stirling) amb l'objectiu d'avaluar-ne el rendiment en ser aplicades a un extens corpus periodístic en noruec. En l'extracció de bigrams, les mesures que obtenen més bons resultats són Z-score, pointwise mutual information i ràtio odds, i quant a trigrams els millors resultats corresponen a informació mútua i ràtio log-likelihood.

Les mesures d'associació lèxica descrites en el present capítol les apliquem de manera experimental en el capítol 6 amb l'objectiu de constatar si són rendibles per a la tasca de validació manual de candidats a terme. Tenint en compte que no totes les mesures d'associació poden ser implementades en diferents tipus d'n-grams, hem seleccionat el tipus d'n-grams que permet una àmplia incorporació de mesures d'associació en el seu processament, tal com recollim en la taula 5.4. En aquest sentit, el processament de candidats bigrams és el que permet un ús més ampli de mesures estadístiques, motiu pel qual en el present capítol descrivim les mesures que poden processar candidats a terme bigrams i en el capítol 6 les implementem en la proposta experimental.

Taula 5.4: Processament de mesures d'associació lèxica.

	Mesures 2-grams	Mesures 3-grams	Mesures 4-grams
Poisson-Stirling	x	x	-
t Student	x	-	-
Prova khi quadrat de Pearson	x	-	-
Ràtio log-likelihood	x	x	x
Ràtio Odds	x	-	-
Informació mútua	x	x	-
Coeficient Dice	x	-	-
c	x	-	-
Coeficient PHI^2	x	-	-
Pointwise mutual information	x	x	-
Freqüència	x	x	x

Les mesures d'associació lèxica compten amb una dimensió lingüística i una dimensió estadística. En la *dimensió lingüística* les mesures es distingeixen per identificar un terme o col·locació com a unitat lingüística sintagmàtica que té una cohesió lèxica (*unithood*) o bé com a unitat lin-

güística relacionada amb els conceptes del seu propi àmbit d'especialitat (*termhood*). Kageura defineix els conceptes *unithood* i *termhood* de la manera següent:

> "*Unithood* refers the degree of strength or stability of syntagmatic combinations and collocations. [...] *Termhood* refers to the degree that a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts." (Kageura i Umino, 1996, p. 11)

Concretament, *unithood* caracteritza unitats lingüístiques complexes formades per paraules que tenen una forta associació (paraules compostes, expressions idiomàtiques i termes complexos); en canvi, *termhood* fa referència al nivell que una unitat lingüística està relacionada amb els conceptes propis d'un àmbit i és pertinent tant per a unitats lingüístiques complexes com per a unitats simples.

Amb relació a la *dimensió estadística*, les mesures d'associació lèxica es basen en principis estadístics i es distingeixen pel seu enfocament metodològic: importància d'associació, força d'associació, teoria de la informació i heurística (Evert i Krenn, 2001, 2004a,b).

Seguidament descriurem les mesures d'associació lèxica a partir de la seva dimensió estadística i lingüística, juntament amb el corresponent enfocament metodològic.

5.2 Mesures d'importància d'associació

Les mesures d'importància d'associació provenen majoritàriament de les proves d'hipòtesis estadístiques. Aquesta aproximació mesura la quantitat d'evidència que hi ha en la mostra observada en contra de la hipòtesi nul·la d'independència com a puntuació d'associació.

Aquest tipus de mesures han estat àmpliament emprades amb uns resultats satisfactoris, encara que els resultats que ofereixen presenten un escull significatiu: una puntuació d'associació alta pot representar un alt nivell d'associació entre els components d'un bigram o bé una gran quantitat d'evidència disponible. Aquest grup de mesures no pot distingir entre aquests dos efectes, motiu pel qual estan esbiaixades a favor de parells amb freqüència alta (llevat de la prova khi quadrat de Pearson) (Evert, 2005).

En el grup de mesures d'importància d'associació hi ha classificades les *mesures de versemblança* (Poisson-Stirling) i també les *proves d'hipòtesis asimptòtiques* (*t* de Student, prova khi quadrat de Pearson, ràtio log-likelihood). Aquestes mesures comparteixen una dimensió estadística i una dimensió lingüística de tipus *unithood*.

5.2.1 Mesures de versemblança

Les mesures de versemblança calculen la probabilitat de la taula de contingència observada (o una part d'aquesta) amb una hipòtesi nul·la de no-associació dels *tokens* que són presents en els bigrams. Totes les mesures de versemblança són bilaterals. La puntuació obtinguda per aquest tipus de mesures pot ser multiplicada per -1 en parells associats negativament a fi d'obtenir una mesura unilateral.

Poisson-Stirling

Poisson-Stirling (Quasthoff i Wolff, 2002) mesura la desviació que hi ha entre les dades observades i el que es podria esperar si *token1* i *token2* fossin independents. Com més alta és la puntuació obtinguda amb aquesta mesura menys proves hi ha a favor de dir que les paraules que formen part del bigram són independents i, per tant, més probabilitat hi ha que el bigram sigui una unitat terminològica.

La mesura Poisson-Stirling fa servir la fórmula de Stirling per a acostar-se al logaritme negatiu de la mesura Poisson-likelihood.

$$Poisson - Stirling = n_{11} \times \left(log \frac{n_{11}}{m_{11}} - 1 \right) \tag{5.1}$$

Els resultats obtinguts amb la mesura Poisson-Stirling són semblants als que s'obtenen amb la ràtio log-likelihood. Des d'un punt de vista computacional, aquesta mesura té un comportament semblant al de la mesura Pointwise mutual information (Kohli, 2006, p. 31).

5.2.2 Proves d'hipòtesis asimptòtiques

Les proves d'hipòtesis asimptòtiques es basen en les distribucions normals i eviten les dificultats numèriques de les proves exactes. Calculen una prova estadística la qual indica fins on arriba la desviació de la taula de contingència observada respecte del valor esperat sota la hipòtesi nul·la. La definició de la prova estadística és clau en el disseny de les proves d'hipòtesis asimptòtiques, ja que determina l'ordre de les possibles taules de contingència, segons la quantitat d'evidència que proporcionin respecte de la hipòtesi nul·la.

Prova *t* de Student

La prova *t* de Student (*Student's t-test*) és una prova asimptòtica que també es coneix com mesura *t-score* o *t-test* (Church i Hanks, 1990). Aquesta

prova determina si l'associació entre dues paraules no és aleatòria calculant el quocient dels valors observats i estimats dividit per l'arrel quadrada del valor de la freqüència observada.

$$t - Student = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \tag{5.2}$$

L'anàlisi de mesures d'associació lèxica feta per Evert (2005) indica que *t* de Student és poc fiable per a la detecció de col·locacions, ja que ofereix resultats extremament conservadors. Ara bé, en altres treballs publicats es constata que aquesta mesura permet obtenir bons resultats en la tasca d'extracció de termes i col·locacions i millorar els resultats obtinguts amb la ràtio log-likelihood i la informació mútua, mesures tradicionalment més ben fonamentades per a dur a terme la tasca d'extracció d'aquest tipus d'unitats (Evert i Krenn, 2001; Krenn i Evert, 2001; Wermter, 2009).

Prova khi quadrat de Pearson

La prova khi quadrat de Pearson (*Pearson's chi-square test*), també anomenada X^2, mesura la desviació entre les dades observades i les dades esperades considerant com a hipòtesi nul·la que *token1* i *token2* presents en un bigram són independents. D'aquesta manera, com més alt és el resultat que s'obté amb aquesta mesura menys evidència hi ha a favor d'afirmar que les paraules que formen part d'un bigram són independents (DeGroot i Schervish, 2002). L'estadística de X^2 suma les diferències entre els valors observats i esperats (taules 5.2 i 5.3), i divideix el resultat pel valor esperat.

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5.3}$$

Ràtio log-likelihood

La ràtio log-likelihood (*log-likelihood ratio*) (Dunning, 1993) mesura quina probabilitat hi ha que *token1* i *token2* apareguin junts per casualitat i que siguin independents l'un de l'altre. Així, com més baixa sigui aquesta probabilitat, més evidència hi ha que el bigram sigui una unitat terminològica. La ràtio log-likelihood té en compte tant les paraules d'un bigram que apareixen juntes com les que apareixen de manera independent. La puntuació que atorga a un bigram és la ràtio entre dues versemblances: la versemblança d'una paraula en un bigram mentre una altra paraula hi és present i la versemblança d'aquesta mateixa paraula en un bigram mentre una altra paraula hi és absent. Si la ràtio és elevada, es constata la dependència estadística.

$$Log - likelihood = -2log\frac{L(O_{11}, C_1, r) \times L(O_{12}, C_2, r)}{L(O_{11}, C_1, r1) \times L(O_{12}, C_2, r2)}$$

$$L(k, n, r) = r^k(1-r)^{n-k}$$

$$r = \frac{R_1}{N}, r_1 = \frac{O_{11}}{C_1}, r_2 = \frac{O_{12}}{C_2} \tag{5.4}$$

La ràtio log-likelihood és una mesura molt usada en extracció de terminologia. Els estudis que s'han fet per a avaluar la rendibilitat de les mesures estadístiques aplicades a l'extracció de termes indiquen que la ràtio log-likelihood ofereix sempre la proporció més alta de termes correctes (Roche *et al.*, 2004), i també és així quan els candidats a terme s'endrecen per freqüència. Els resultats experimentals que han dut a terme Daille (1994) i Evert i Krenn (2001) constaten que aquesta ràtio assoleix la precisió més alta en comparació amb altres mesures.

Log-likelihood es diferencia d'altres proves estadístiques pel fet que totes les cel·les de la taula de contingència són tingudes en compte a l'hora de

calcular els resultats; en canvi, la prova *t* de Student i informació mútua només tenen en compte la freqüència observada (O_{11}) i la freqüència esperada (E_{11}). Una altra diferència respecte altres mesures és el fet que sigui una prova bilateral, cosa que representa que els valors alts indiquen un alt nivell d'associació entre els *tokens* que formen part del bigram, sigui en positiu o negatiu (Wermter, 2009).

5.3 Mesures de força d'associació

Les mesures descrites en aquest apartat calculen el coeficient de força d'associació de les dades observades, és a dir, la puntuació d'associació calculada s'estima a partir del nivell d'associació que hi ha entre els *tokens* que formen part d'un bigram. Així mateix, aquestes mesures proporcionen un enfocament diferent de les mesures d'importància d'associació, que se centren en el nivell d'associació.

En aquest grup de mesures hi ha classificades les d'*estimació puntual de força d'associació*, que són la ràtio odds, la informació mútua, el coeficient Dice i el coeficient Jaccard. Totes aquestes mesures comparteixen una dimensió estadística i una dimensió lingüística de tipus *unithood*.

5.3.1 Estimació puntual de força d'associació

Les estimacions puntuals de força d'associació com ara les estimacions *maximum-likelihood* tenen com a inconvenient que el valor d'aquests coeficients no pot determinar del tot la distribució del mostreig; per tant, qualsevol valor hipotetitzat està en consonància amb una àmplia varietat d'estimacions de les dades observades. La solució que s'aplica en aquest cas és fer servir el valor d'estimació més alt dins un rang. Per aquest motiu, l'estimació del coeficient de força d'associació es calcula a partir de l'estimació directa del conjunt de paràmetres τ_{ij}, per als quals la probabilitat total de la taula de contingència observada assumeix el seu màxim global.

Ràtio odds

La ràtio odds o oportunitat relativa (θ) (*odds ratio*) (Everitt, 1992) calcula la ràtio entre el nombre de vegades que els *tokens* d'un bigram apareixen junts i el nombre de vegades que els *tokens* apareixen individualment. És el resultat de multiplicar en diagonal els resultats de la taula de contingència 2x2.

$$odds = \frac{n_{11} \times n_{22}}{n_{21} \times n_{12}} \tag{5.5}$$

La ràtio odds aplicada a la identificació de termes dóna un valor molt alt als termes que apareixen en un document, encara que la freqüència d'aquests termes sigui molt baixa, ja que aquesta ràtio tendeix a sobreestimar les freqüències baixes (Tsay i Wang, 1999).

Informació mútua

La informació mútua (*mutual information*) (Fano, 1961) mesura la dependència entre dues variables o conjunts de variables. La informació mútua I(X;Y) és la reducció de la incertesa d'una variable aleatòria (*x*) pel fet de conèixer el valor d'una altra variable aleatòria (*y*).

Fano inicialment va definir la informació mútua entre esdeveniments particulars *x'* i *y'*, que en el nostre cas corresponen a l'ocurrència de paraules concretes. Si dos punts (paraules), *x* i *y*, tenen les probabilitats *P(x)* i *P(y)*, la seva informació mútua, *I(x y)*, es defineix de la manera següent:

$$I(xy) = log_2 \frac{P(xy)}{P(x)P(y)} \tag{5.6}$$

La informació mútua compara la probabilitat d'observar *x* i *y* juntes (probabilitat conjunta) amb les probabilitats d'observar *x* i *y* independents (probabilitat aleatòria). Si hi ha una veritable associació entre *x* i *y*, la probabilitat conjunta *P(x,y)* és més alta que l'aleatòria *P(x) P(y)* i, en conseqüència, $I(x, y) \gg 0$. Si no hi ha cap tipus de relació entre *x* i *y*,

llavors $P(x, y) \approx P(x)P(y)$ i, així, $I(x, y) \approx 0$. Si *x* i *y* es troben en una distribució complementària, llavors *P(x,y)* és molt menys que *P(x) P(y)*, forçant $I(x, y) \ll 0$ (Church i Hanks, 1990).

Aquesta mesura és àmpliament usada per al processament del llenguatge natural estadístic, com ara en la classificació de paraules (*word clustering*) o la desambiguació semàntica (*word sense disambiguation*).

Coeficient Dice

El coeficient Dice (*Dice coefficient*) (Dice, 1945; Smadja *et al.*, 1996) identifica bigrams que tenen un alt nivell de cohesió lèxica. Si el coeficient Dice atorga un valor alt a un bigram significa que els *tokens* que en formen part no apareixen junts per casualitat i, per tant, el bigram en qüestió és rellevant. En aquest sentit, el coeficient Dice és alt quan els *tokens* presents en els bigrams ocorren junts amb més freqüència que no pas individualment (McInnes, 2004; Kohli, 2006).

$$Dice = \frac{2 \times n_{11}}{n_{1p} + n_{p1}} \tag{5.7}$$

Coeficient Jaccard

El coeficient Jaccard (*Jaccard coefficient*) (Jaccard, 1901; Dunning, 1998) calcula el nombre de vegades que com a mínim un dels *tokens* que formen part d'un bigram es troba en posició correcta. El coeficient Jaccard es calcula de la manera següent:

$$Jaccard = \frac{n_{11}}{n_{11} + n_{12} + n_{21}} \tag{5.8}$$

El coeficient Jaccard és semblant al coeficient Dice, per aquest motiu també pot ser calculat aplicant la transformació del coeficient Dice:

$$Jaccard = \frac{Dice}{2 - Dice} \tag{5.9}$$

Aquest coeficient és molt usat en l'àmbit de la recuperació d'informació com a mesura d'associació. Es fa servir per a mesurar el nivell d'associació entre dues variables (Kohli, 2006).

Coeficient PHI^2

El coeficient PHI^2 (PHI^2 *coefficient*) (Church i Gale, 1991) mesura el nivell d'associació entre dues variables binàries. En els bigrams, aquestes variables corresponen als *tokens* que en formen part, indiquen si un determinat *token* hi és present o no i en quina posició ho fa (Kohli, 2006). Church va fer servir el coeficient PHI^2 per a identificar bigrams en corpus textuals. Aquest coeficient es calcula de la manera següent:

$$phi^2 = \frac{\left(n_{11} \times n_{22} - n_{12} \times n_{21}\right)^2}{n_{1p} \times n_{p1} \times n_{2p} \times n_{p2}} \quad (5.10)$$

El resultat d'aquest coeficient se situa entre el límit de 0 i 1, el qual indica el nivell de força d'associació que hi ha entre dos *tokens*. Aquest coeficient tendeix a afavorir combinacions de *tokens* que siguin freqüents.

El coeficient PHI^2 ha estat usat per a localitzar concordances en textos paral·lels i també s'ha fet servir en la tasca d'identificar candidats a terme en models híbrids (lingüístics i estadístics) d'extracció automàtica de terminologia (Daille, 1994, 1997).

5.4 Mesures de la teoria de la informació

Les mesures provinents de la teoria de la informació es basen en conceptes teòrics com ara *entropia*, *cross-entropia* i *informació mútua*. Les mesures d'associació d'aquest grup quantifiquen la no-homogeneïtat de la taula de contingència observada comparada amb les freqüències esperades de la taula de contingència.

En el grup de mesures extretes de la teoria de la informació hi ha classificada la mesura pointwise mutual information, que és una variant d'informació mútua. Aquesta mesura compta amb una dimensió estadística i una dimensió lingüística, de tipus *unithood*.

5.4.1 Pointwise mutual information

Pointwise mutual information (PMI) (Church i Hanks, 1990) mesura el nivell de coincidència que hi ha entre *tokens*. Dit d'una altra manera, mesura el nivell d'informació que ofereix la presència d'un *token* en una determinada posició respecte d'un altre *token*. Així, Pointwise mutual information calcula com s'incrementa la quantitat informació sabent que un determinat *token* se situa en primera posició en un bigram (*token1*) i que va seguit d'un altre *token* (*token2*).

Pointwise mutual information es defineix com el logaritme de la desviació entre la freqüència observada del bigram (*n11*) i la probabilitat que aquest bigram sigui independent (*m11*).

$$PMI = log\left(\frac{n_{11}}{m_{11}}\right) \tag{5.11}$$

S'han publicat diferents definicions d'informació mútua, com es desprèn dels treballs de Fano (1961) i Cover i Thomas (2006), i en la taula 5.5 mostrem quina correlació tenen (Manning i Schütze, 2003).

Taula 5.5: Definicions d'informació mútua.

Símbol	Definició	Ús actual	Fano (1961)
$I(x,y)$	$log\frac{p(x,y)}{p(x)p(y)}$	pointwise mutual information	mutual information
$I(X;Y)$	$Elog\frac{p(X,Y)}{p(X)p(Y)}$	mutual information	average MI expectation of MI

Pointwise mutual information va ser una de les primeres mesures a ser introduïda en la lingüística computacional amb l'objectiu de localitzar els candidats a terme interessants d'un text, ja que mesura la correlació de paraules en un bigram. Aquesta mesura s'aplica en la tasca de detecció de sinònims fent servir la freqüència de les coocurrències obtingudes a partir de les consultes fetes a un motor de cerca (Turney, 2001). Així mateix, la mesura Pointwise mutual information permet l'extracció de terminologia d'un corpus. I la seva variant, MI^3 (Daille *et al.*, 1998), incrementa la puntuació que obtenen els candidats a terme extrets d'un corpus.

$$MI^3(xy) = log_2 \frac{P(xy)^3}{P(x)P(y)} \tag{5.12}$$

5.5 Mesures heurístiques

Les mesures heurístiques combinen valors de mostra que són considerats bons indicadors d'associació (positiva). Es basen en assumpcions empíriques i intuïtives, a diferència dels altres tres grups de mesures, que tenen un fort rerefons estadístic. En el grup de mesures heurístiques hi ha classificada la freqüència, la qual compta amb una dimensió estadística i una dimensió lingüística, de tipus *termhood*.

5.5.1 Freqüència

La freqüència és la mesura d'associació més simple basada en la coocurrència de *tokens*. L'ús d'aquesta mesura ve motivat per l'assumpció que parells de *tokens* associats apareixen en general amb més freqüència que les combinacions arbitràries, fet que està relacionat amb el criteri de recurrència (Firth, 1957). En una avaluació empírica de les mesures d'associació, la freqüència és utilitzada com a punt de referència amb la qual altres mesures més complexes són comparades (Evert i Krenn, 2001).

$$Freqncia = O_{11} \tag{5.13}$$

La freqüència no deriva de cap principi teòric estadístic, sinó d'una simple assumpció que diu que "una expressió freqüent indica un concepte important d'un àmbit específic i, en conseqüència, hauria de tenir una posició elevada en el rang de candidats a terme". L'objecció més important que es fa a l'ús de la freqüència com a mesura per al reconeixement de termes és que no té en consideració el nivell o grau d'associació (*unithood*) de les paraules que formen unitats multiparaula (Velardi *et al.*, 2001). D'aquesta manera, expressions molt freqüents són considerades bons candidats encara que no siguin termes. Així, doncs, per a tenir en compte el nivell d'associació dels termes fent servir la freqüència, cal usar filtres lingüístics que puguin descartar candidats que no tinguin unes determinades propietats sintàctiques o morfològiques (Justeson i Katz, 1995).

Diferents estudis basats en extracció de col·locacions (Evert i Krenn, 2001; Krenn i Evert, 2001) i en extracció de terminologia (Daille, 1997) indiquen que el rendiment que assoleix la mesura de freqüència en la detecció de candidats a terme es troba al mateix nivell d'altres mesures estadístiques, com ara *t* de Student o log-likelihood (Wermter, 2009).

Així mateix, és rellevant l'ús de la *freqüència relativa* per a la recuperació de termes. Aquesta mesura va ser definida per primera vegada per Edmunson i Wyllys de la manera següent:

> "[...] it would seem natural to regard the contrast between the word's relative frequency f within the document and its relative frequency r in general use... as a more revealing indication of the word's value in indicating the subject matter of a document. Such a contrast can be represented by the ratio f/r [...]." (Edmundson i Wyllys, 1961, p. 227)

Tenint en compte que els termes acostumen a aparèixer més sovint en textos especialitzats del seu propi domini temàtic que no pas en textos de llengua general, Ahmad *et al.* (1994) i Damerau (1993) indiquen que les unitats terminològiques poden ser identificades comparant la freqüència

relativa de les paraules que apareixen en un text especialitzat amb la freqüència relativa que tenen en un corpus més ampli i temàticament variat. D'aquesta manera, les paraules que apareixen de manera significativa més vegades en el corpus especialitzat del que s'esperaria tenint en compte la seva freqüència relativa en el corpus de referència, són les que han de ser extretes com a unitats terminològiques. L'ús de la freqüència relativa és determinant per a la identificació dels termes monoparaula que apareixen en els corpus d'especialitat.

5.6 Recapitulació

En el present capítol hem descrit les mesures d'associació lèxica que implementem en la nostra proposta experimental i que presentem en el capítol 6. D'aquesta descripció s'observa que la puntuació d'associació que calculen les mesures per cada candidat a terme pot ser considerada per a estimar la importància d'associació entre els elements que formen part d'un candidat a terme, per a obtenir el rànquing dels candidats a terme extrets d'un corpus d'especialitat o bé per a establir el rànquing dels candidats que estiguin formats per uns determinats *tokens*.

Les mesures d'associació lèxica que hem presentat en aquest capítol compten amb una dimensió estadística i una dimensió lingüística. En la dimensió estadística les mesures atribueixen un valor de rang als candidats a terme tenint en compte la importància d'associació i la força d'associació que s'estableix entre els *tokens* que formen part del bigram, i també la teoria de la informació i l'heurística. En la dimensió lingüística el valor de rang que obté un candidat a terme té en compte el nivell de cohesió lèxica dels *tokens* que formen part del candidat *unithood* i el grau de pertinença d'un candidat a un domini d'especialitat *termhood*. La descripció de les mesures l'hem organitzada a partir de les dimensions estadística i lingüística, i el corresponent enfocament metodològic (Evert i Krenn, 2004a; Pazienza *et al.*, 2005).

La classificació i descripció de les mesures d'associació lèxica ens ha permès observar que la rendibilitat que ofereixen s'ajusta a l'objectiu de treball plantejat en el capítol 6, tal com indiquen els estudis que comparen la rendibilitat de les mesures. Així mateix, la principal conclusió que es desprèn d'aquests estudis és que hi ha importants diferències en els resultats que s'obtenen de les mesures d'associació per a la tasca d'extracció de candidats d'un corpus amb finalitat terminològica.

El conjunt de mesures descrites en el present capítol (taula 5.6) són avaluades en la proposta experimental del capítol 6, amb l'objectiu identificar quina mesura o conjunt de mesures permet identificar un major nombre de termes d'un corpus d'especialitat per a agilitar la tasca final de validació de candidats que duu a terme un especialista.

Taula 5.6: Dimensions estadística i lingüística (*unithood* i *termhood*).

		Dimensió estadística			
Dimensió lingüística	Unithood	**Importància associació**	**Força associació**	**Teoria informació**	**Mesures heurístiques**
		Poisson-Stirling	Odds Inf. mútua Dice Jaccard PHI^2	Pointwise MI	
		t Student X^2 Log-likelihood			
	Termhood				Freqüència

Capítol 6

VALIDACIÓ DE TERMES I MESURES D'ASSOCIACIÓ LÈXICA

En el present capítol plantegem una aproximació experimental basada en mesures d'associació lèxica que han estat tradicionalment aplicades a l'extracció automàtica de terminologia (capítol 5). Aquesta aproximació experimental té com a objectiu constatar si les mesures d'associació lèxica permeten reduir en temps i cost la tasca de validació manual de candidats a terme que fan els especialistes d'un àmbit d'especialitat (Merkel i Foo, 2007). La validació manual dels resultats obtinguts en la fase d'extracció automàtica de termes permet destriar quins candidats són termes propis d'un àmbit d'especialitat; ara bé, té l'inconvenient que és molt lenta i costosa, demana un bon coneixement de l'àmbit d'especialitat i molta experiència per part dels especialistes (Foo, 2011). Per tal d'observar la rendibilitat de les mesures d'associació lèxica per a la validació de termes, avaluarem el valor de rang que assignen les mesures a cada un dels candidats que s'obtenen en un procés d'extracció automàtica de termes, contrastarem els resultats obtinguts amb el càlcul estadístic de freqüència i el mètode TSR i analitzarem els avantatges que ofereix la implementació de mesures a la tasca d'extracció automàtica de termes.

6.1 Ús de mesures d'associació lèxica

La incorporació de mesures d'associació lèxica a la tasca d'extracció automàtica de termes presents en corpus especialitzats té com a objectiu explorar quina capacitat tenen aquestes mesures de situar en uns determinats rangs els termes que hi ha en una llista de candidats per a reduir en temps i cost el procés de validació manual dels termes.

Per tal de fer efectiva aquesta proposta experimental, avaluem la rendibilitat que poden oferir les mesures d'associació lèxica en ser implementades en la fase final d'un procés d'extracció automàtica de termes. Més concretament, incorporem les mesures per a processar els resultats obtinguts amb el mètode freqüència i el mètode TSR descrits en l'apartat 4.3. Així mateix, avaluem la rendibilitat de les mesures tenint en compte els candidats a terme bigrams extrets automàticament dels corpus especialitzats que s'empren en la present recerca. Centrem l'anàlisi en candidats bigrams pel fet que els termes formats per dos *tokens* tenen una presència majoritària en els corpus especialitzats (Pecina i Schlesinger, 2006) i, alhora, permeten una àmplia aplicació de mètodes estadístics, escalabilitat que es veu reduïda a mesura que els *n-grams* seleccionats són d'ordre major (Nakagawa i Mori, 2002; Pecina, 2010). En la taula 6.1 aportem una mostra de la distribució dels termes en els corpus especialitzats que emprem en la present recerca.

Taula 6.1: Distribució dels termes en els corpus especialitzats.

Corpus	2-grams	3-grams	4-grams	5-grams	6-grams
Economia espanyol	522	299			
Medicina espanyol	665	82			
Serveis socials espanyol	119	38	26	24	9
Serveis socials català	127	48	29	22	14

El processament dels resultats obtinguts dels mètodes de freqüència i TSR fent ús de les mesures d'associació lèxica és dut a terme amb l'eina d'anàlisi estadística Ngram Statistics Package (Text-NSP). Aquesta eina consta d'un conjunt d'utilitats que ajuden a analitzar els *n-grams* presents en un corpus fent servir tests estàndards d'associació. A partir de la llista d'*n-grams* d'un corpus i una mesura d'associació lèxica, l'eina calcula la puntuació que té cada *n-gram*, i aquest és situat en un rang determinat. La llista d'*n-grams* final queda endreçada per ordre descendent de rang. La puntuació estadística que es calcula per *n-gram* serveix per a decidir si hi ha prou evidència o no per a rebutjar la hipòtesi nul·la per cada *n-gram*, és a dir, si l'*n-gram* és una unitat terminològica o no ho és. El nombre de tests estàndards d'associació que incorpora l'eina Text-NSP permet fer una exploració en profunditat del nivell de rendibilitat que poden oferir les mesures estadístiques per a la tasca de validació manual final de termes.

Seguidament descrivim els resultats que s'obtenen en implementar les mesures d'associació lèxica a la llista de candidats a terme extrets amb el mètode de freqüència i el mètode TSR.

6.1.1 Resultats de les mesures amb el mètode freqüència

Les mesures d'associació lèxica aplicades al mètode freqüència processen els candidats a terme extrets dels corpus especialitzats (apartat 4.1.1), prèviament filtrats per paraules buides i endreçats per freqüència amb l'eina Text-NSP. Aquest procés d'extracció de candidats és un dels més habituals en la tasca d'extracció automàtica de terminologia.

El processament dels candidats a terme endreçats per freqüència amb les mesures permet atribuir un valor de rang a cada un dels candidats. Cada mesura d'associació lèxica calcula un valor de rang diferent per cada un dels candidats extrets i, per tant, el resultat inicial obtingut amb el mètode freqüència és reendreçat per cada mesura tenint en compte la importància d'associació, la força d'associació, la informació mútua i les heurístiques.

Els candidats a terme endreçats per freqüència i els candidats reendreçats segons cada mesura d'associació lèxica són comparats amb els termes seleccionats manualment dels corpus especialitzats per a poder analitzar en quin rang o posició de freqüència queden situats els termes extrets automàticament. Amb aquest contrast volem determinar el rendiment de les mesures d'associació lèxica aplicades al mètode estadístic de freqüència.

En les taules 6.2 i 6.3 recollim la distribució dels termes per corpus a partir de posicions que hem seleccionat. En cada taula comparem el nombre de termes identificat amb el mètode freqüència i les mesures d'associació lèxica i en destaquem el valor més alt.

Taula 6.2: Distribució de rangs amb el mètode freqüència (I).

Corpus economia espanyol JRC (UE)											
Top	Freq	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
100	**40**	9	31	20	9	1	32	31	**37**	9	9
500	**86**	59	72	55	55	42	73	72	**77**	55	59
800	**110**	102	103	100	103	95	103	103	**104**	103	102
Corpus economia anglès JRC (UE)											
Top	Freq	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
100	**34**	2	26	13	2	0	27	26	**34**	2	2
500	**77**	24	53	19	23	13	53	54	**72**	23	24
1000	**90**	48	70	43	52	37	73	70	**84**	52	48
1800	**125**	119	120	114	119	114	119	120	**121**	119	119
Corpus economia francès JRC (UE)											
Top	Freq	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
100	**24**	9	**24**	16	9	3	**24**	**24**	**25**	9	9
500	**53**	40	47	37	38	30	48	47	**51**	38	40
800	69	69	71	68	69	62	**72**	69	**73**	69	69

Taula 6.3: Distribució de rangs amb el mètode freqüència (II).

Corpus economia espanyol IULA (UPF)											
Top	Freq	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
100	**50**	7	42	12	7	5	43	42	**47**	7	7
500	**126**	34	90	26	34	22	91	91	**117**	34	34
1000	**163**	87	120	75	84	53	121	120	**135**	84	87
2000	**236**	201	210	203	207	200	209	210	**212**	207	201
Corpus serveis socials català Termcat											
Top	Freq	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
100	**15**	1	11	5	1	0	11	11	**15**	1	1
500	**28**	15	**28**	14	16	8	**28**	**28**	**30**	16	15
800	32	22	**36**	20	24	17	**35**	**36**	34	24	22
1000	36	29	**38**	29	30	22	**37**	**38**	**37**	30	29
Corpus serveis socials espanyol Termcat											
Top	Freq	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
100	**18**	1	14	6	1	1	14	14	**19**	1	1
500	**37**	20	33	19	17	9	32	33	**36**	17	20
800	**45**	27	39	26	28	27	37	38	**41**	28	27
1000	**51**	35	**47**	32	37	30	46	**47**	**47**	37	35
Corpus medicina espanyol IULA (UPF)											
Top	Freq	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
100	**17**	3	13	5	3	1	13	13	**15**	3	3
500	**47**	10	42	15	10	9	39	42	**45**	10	10
1000	**70**	29	58	27	28	22	60	57	**70**	27	29
3000	**120**	81	112	76	83	66	111	108	**116**	82	81
5000	**151**	127	145	129	135	117	146	148	**149**	135	127

Els resultats que hem obtingut dels corpus especialitzats permeten observar que la freqüència i la mesura *t* de Student situen un major nombre de termes en les diferents posicions seleccionades i en sis dels set corpus analitzats. Les mesures ràtio log-likelihood, Poisson-Stirling i informació mútua destaquen per obtenir més bons resultats que no pas la freqüència i la mesura *t* de Student en un dels set corpus.

En aquest sentit, les dades constaten que l'endreçament de candidats per freqüència situa un nombre significatiu de termes en les posicions inicials dels resultats, en comparació amb les mesures d'associació lèxica avaluades (Krenn i Evert, 2001; Wermter i Hahn, 2006; Wermter, 2009). Amb tot, considerem que l'aportació que fan les mesures d'associació amb l'endreçament dels candidats per mitjà dels rangs és consolidar les posicions que assoleixen els termes amb el mètode de freqüència i, en conseqüència, la rellevància que tenen en el conjunt dels candidats. El valor de rang és un indicador que aporten les mesures i que determina el grau d'associació que s'estableix entre les paraules que formen part dels candidats; així, com més alt és el rang més alt és el nivell d'associació entre les paraules que configuren els candidats i major probabilitat hi ha que la unitat sigui terminològica.

La informació de freqüència i rang que s'atribueix conjuntament a un candidat a terme facilita la identificació dels termes pel fet que tant els candidats amb més presència al corpus com els que tenen un nivell més alt de rang quedaran situats en les posicions inicials dels resultats. Si es dóna el cas que conflueixen aquests dos trets en una mateixa unitat, augmenta encara més la probabilitat que aquesta unitat sigui terminològica.

En definitiva, els resultats obtinguts mostren que la implementació de la mesura *t* de Student en el procés d'extracció automàtica de candidats endreçats per freqüència permet consolidar les posicions en les quals queden situats els termes i facilitar la seva identificació en la tasca de validació manual final de candidats.

6.1.2 Resultats de les mesures amb el mètode TSR

La implementació de mesures d'associació lèxica en el procés d'extracció automàtica de termes amb el mètode TSR es produeix en el moment en què els candidats dels corpus especialitzats són filtrats per *tokens* terminològics (apartat 4.2). Els resultats obtinguts del filtratge amb el mètode TSR són processats per l'eina Text-NSP amb cada una de les mesures estadístiques descrites en el capítol 5. D'aquest processament s'obté una llista de candidats a terme endreçats per ordre de rang, el qual és calculat per cada mesura estadística. El rang que assignen les mesures als candidats indica el grau de cohesió lèxica que té un candidat i, per tant, a major grau de cohesió més probabilitat hi ha que el candidat sigui un terme.

Els candidats endreçats per rang són contrastats amb els termes seleccionats manualment dels corpus especialitzats que s'empren en la present recerca per tal d'establir una correlació entre nombre de termes i nombre de candidats per rang coincidents amb aquests termes. D'aquesta manera, comprovem quina mesura situa un major nombre de termes en les posicions inicials dels resultats i, per tant, facilita la tasca de validació manual final dels candidats.

Les taules 6.4 i 6.5 mostren la distribució per posicions dels termes endreçats amb el mètode TSR i els rangs de les diferents mesures d'associació lèxica en els corpus especialitzats analitzats. De cada corpus en destaquem els valors més alts.

Taula 6.4: Distribució de termes per rangs amb el mètode TSR (I).

Corpus economia espanyol JRC (UE)											
Top	TSR	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
10	**9**	5	**9**	**8**	5	2	**9**	**9**	**9**	5	5
50	**30**	24	26	23	24	20	26	26	**28**	24	24
100	**43**	**42**	41	39	41	37	41	41	**42**	41	**42**
150	**55**	**55**	**55**	**54**	**55**	**54**	**55**	**55**	**55**	**55**	**55**
Corpus economia anglès JRC (UE)											
Top	TSR	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
10	**8**	**7**	6	6	**7**	4	6	6	**7**	**7**	**7**
50	**28**	18	25	15	17	10	25	25	**27**	17	18
100	**41**	32	35	26	27	20	36	35	**40**	27	32
200	54	52	**56**	47	51	44	53	**56**	**55**	51	52
350	**81**	77	77	77	77	77	**78**	77	77	77	77
Corpus economia francès JRC (UE)											
Top	TSR	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
10	**8**	6	**7**	**7**	6	6	**7**	**7**	**8**	6	6
50	22	**25**	**24**	23	**24**	22	**24**	**24**	**24**	**24**	**25**
90	**39**	**39**	**38**	**39**	**38**	**39**	**38**	**38**	**39**	**38**	**39**
Corpus economia espanyol IULA (UPF)											
Top	TSR	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
10	**10**	6	**9**	5	5	4	**9**	**9**	**10**	5	5
50	**43**	26	40	20	23	13	39	40	**42**	23	26
100	**68**	48	**59**	43	48	26	58	**59**	**68**	48	48
150	**88**	62	71	55	65	48	72	71	**81**	65	62
200	**98**	76	90	76	80	62	91	90	**96**	80	76
300	**123**	112	**116**	110	115	106	115	**116**	114	115	112
400	**149**	**149**	**149**	**148**	**148**	**148**	**148**	**148**	**149**	**148**	**149**

Taula 6.5: Distribució de termes per rangs amb el mètode TSR (II).

Corpus serveis socials català Termcat											
Top	TSR	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
10	**8**	3	6	3	3	0	**7**	6	6	3	3
50	**17**	14	**15**	14	14	13	14	**15**	**15**	14	14
75	**25**	**25**	24	**26**	24	**26**	**26**	24	**26**	24	**25**
Corpus serveis socials espanyol Termcat											
Top	TSR	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
10	**7**	4	**7**	4	4	1	5	**6**	**6**	4	4
50	**22**	17	**21**	18	19	18	18	**21**	20	19	17
80	**34**	32	**33**	**34**	**33**	**34**	32	**33**	**34**	**33**	31
Corpus medicina espanyol IULA (UPF)											
Top	TSR	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
10	**8**	3	**6**	2	4	3	**6**	**6**	**8**	4	3
50	**26**	18	**24**	16	17	15	23	**24**	**26**	17	18
100	**40**	35	**42**	33	33	22	39	**42**	39	33	35
200	**68**	56	61	47	61	45	62	61	**67**	61	56
300	**87**	74	86	68	70	61	80	85	**89**	70	74
400	**103**	93	96	87	90	83	94	97	**100**	90	93
600	131	131	134	125	**135**	122	133	**134**	132	**135**	131
750	155	151	**157**	155	**157**	155	155	**158**	**156**	**157**	151

Els resultats obtinguts constaten que *t* de Student és la mesura que situa un major nombre de termes en les diferents posicions analitzades. Les mesures ràtio log-likelihood i informació mútua també situen un nombre significatiu de termes en posicions inicials, però en menor nombre que la mesura *t* de Student. El coeficient Dice, la ràtio odds, el coeficient PHI^2, pointwise mutual information, poisson-Stirling, la prova khi quadrat i el coeficient Jaccard situen un nombre reduït de termes en posicions inicials. Convé notar que com més alta és la posició en què se situen els candidats

a terme més igualtat hi ha en el nombre de termes identificats per cada mesura.

En aquest sentit, la incorporació de mesures d'associació lèxica en el procés d'extracció automàtica de termes amb el mètode TSR permet constatar que *t* de Student és la mesura que endreça en major nombre i en les posicions inicials dels resultats els termes presents en tots els corpus especialitzats. En conseqüència, aquesta mesura és la que fa possible en ambdós mètodes d'extracció de candidats (mètode freqüència i mètode TSR) consolidar les posicions en què s'han situat inicialment els termes durant els processos d'extracció i facilitar la identificació dels termes per la posició inicial de rang amb què compten.

Així mateix, si analitzem les mesures des del punt de vista de la rendibilitat que poden oferir en la tasca de validació manual de candidats a terme d'un corpus, també convé observar quina mesura situa un major nombre de termes en el menor nombre de candidats possible. Per a fer-ho, de cada mesura hem comprovat el nombre de candidats que cal revisar per a poder identificar el 25%, el 50% i el 75% dels termes presents en els corpus. D'aquesta manera, obtenim el nombre de candidats que s'ha de revisar manualment en cada una d'aquestes tres franges percentuals i el percentatge que representa respecte del total de candidats. En l'annex A hi ha recollides les taules A.1, A.2, A.3, A.4 i A.5, en les quals indiquem quin és el nombre de termes i candidats que conté cada corpus i també el nombre de candidats que cal revisar manualment equivalent a un 25%, 50% i 75% dels resultats. En cada franja percentual destaquem el nombre més alt i més baix de candidats a terme que s'ha de revisar manualment per a arribar a identificar els termes que hi són presents.

L'anàlisi de les dades recollides en les taules A.1, A.2, A.3, A.4 i A.5 per tal d'identificar el nombre de candidats que ha de ser revisat manualment de cada mesura a fi d'extreure el 25%, el 50% i el 75% dels termes presents en els corpus especialitzats, permet establir les correlacions que presentem a continuació:

- El 25% dels termes presents en els corpus especialitzats queden situats entre el 9,5% i el 10,3% dels candidats a terme, que corresponen als rangs més alts en què els termes queden endreçats per les mesures ràtio log-likelihood, Poisson-Stirling, informació mútua i també *t* de Student. Aquests percentatges corresponen als corpus d'economia espanyol de l'IULA i economia anglès de JRC.

- El 50% dels termes presents en els corpus especialitzats queden situats entre el 28,4% i el 32,8% dels candidats, que equivalen als rangs més alts en què els termes són situats per les mesures coeficient Dice, ràtio log-likelihood, informació mútua i *t* de Student. Aquest resultat correspon als corpus d'economia espanyol i anglès de JRC i medicina espanyol de l'IULA.

- El 75% dels termes presents en els corpus queden situats entre el 54,6% i el 55,8% dels candidats per part de les mesures ràtio log-likelihood, informació mútua i *t* de Student. Aquest resultat correspon al corpus de serveis socials espanyol del Termcat.

A part de les correlacions que acabem de descriure i que incideixen en els rangs més alts en què se situen els termes, convé observar en quins percentatges se situen els termes dels corpus analitzats per tal d'identificar el nombre de candidats que ha de ser revisat manualment. La taula 6.6 mostra els percentatges en els quals queden situats els termes en cada mesura d'associació lèxica.

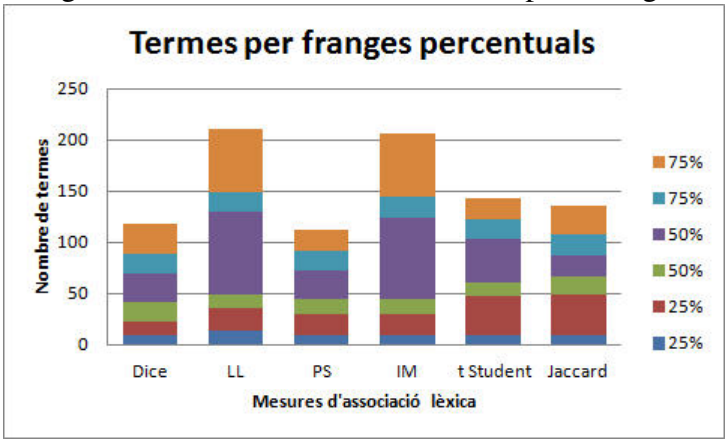
Taula 6.6: Distribució dels candidats a terme per franja percentual.

%	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* Stud.	X^2	Jacc.
25	12,1-25,5	**9,5-17,8**	14,5-24,4	15,4-27,9	16,6-35	**10,1-16,2**	**9,5-17,4**	**10,3-17,4**	15,4-27,9	14,4-25,5
50	**30,5-59,3**	**32,8-61,2**	38,8-61,2	36,3-58,7	42,7-63,7	34,3-58,7	**32,8-61,2**	**28,4-58,7**	36,3-58,7	**32,2-59,3**
75	61,4-81,2	**55,8-74,5**	61,6-77,5	59,3-75	69,7-76,7	63,9-73,7	**55,8-74,7**	**54,6-75**	59,3-75	61,4-81,2

El coeficient Dice, la ràtio log-likelihood, la mesura Poisson-Stirling, la mesura informació mútua, la mesura *t* de Student i el coeficient Jaccard són les sis mesures que situen els termes en els rangs més alts per franges percentuals. Així mateix, s'observa que log-likelihood, informació mútua i *t* de Student són les úniques mesures que situen els termes del conjunt de corpus analitzats en els rangs més elevats, i ho fan en les tres franges percentuals. Concretament, situen el 25% de termes dels corpus entre el 9,5% i el 17,8% de candidats, el 50% de termes els situen entre el 28,4% i el 61,2% de candidats i el 75% de termes entre el 54,6% i el 75% de candidats respectivament. Les mesures Dice, Poisson-Stirling i Jaccard tenen una rellevància baixa en el global de les dades analitzades pel fet de situar els termes en els rangs més alts en una sola franja percentual.

Si observem gràficament la distribució dels termes en les tres franges percentuals per part de les mesures que obtenen millors resultats en la taula 6.6, constatem que les mesures log-likelihood i informació mútua, seguides de *t* de Student, són les que situen un major nombre de termes en cada franja percentual (figura 6.1).

Figura 6.1: Distribució dels termes en percentatges.



Completem l’anàlisi quantitativa de les dades aportades amb exemples de termes que extreuen les mesures d’associació lèxica que assoleixen una major rendibilitat de resultats en l’extracció. Concretament seleccionem una mostra dels primers candidats corresponents a les mesures log-likelihood, informació mútua i *t* de Student, les quals assoleixen els millors resultats en el reendreçament de termes en la totalitat dels candidats en cinc dels set corpus analitzats per mesura.

Per cada una d’aquestes tres mesures incloem exemples dels corpus en els quals situen un major nombre de termes en el menor nombre de candidats en la franja del 25% dels resultats analitzada, i, en conseqüència, redueixen en temps la tasca manual final de revisió dels candidats. Les llistes d’exemple pertanyen als dominis de l’economia i la medicina en espanyol, anglès i francès, i mostren un predomini de termes per sobre dels candidats en tots els casos. S’observa una gran similitud en l’ordre en què queden endreçats els termes d’un mateix corpus en mesures diferents, cosa que indica una homogeneïtat en l’establiment de l’ordre de rang dels candidats en aquestes tres mesures. En les taules 6.7, 6.8, 6.9 hi ha distribuïts els candidats, i els que són termes estan marcats en cursiva.

Taula 6.7: Distribució de termes per corpus i mesures (I).

Log-likelihood		
Econ. ang. JRC	Econ. esp. IULA	Med. esp. IULA
euro area	*política económica*	*asma bronquial*
Lisbon strategy	*equilibrio general*	*sexo masculino*
internal market	*precio neto*	*función pulmonar*
human capital	*entorno económico*	*sexo femenino*
framework conditions	*elección social*	*hiperreactividad bronquial*
public procurement	*sector oligopolístico*	grupo control
social partners	*equilibrio parcial*	consulta externa
productivity growth	*reforma fiscal*	capacidad vital
sustainable development	informacionalmente eficiente	*diabetes mellitus*
industrial base	*competencia perfecta*	*flujo espiratorio*
competitive advantages	*coste marginal*	*sangre periférica*
structural reforms	*problema económico*	*hipertensión arterial*
energy efficiency	*renta disponible*	*volumen espiratorio*
economic recovery	*margen unitario*	*actividad fagocítica*
macroeconomic policies	Comunidad Económica	pulmonar obstructiva
labour market	*sector público*	*insuficiencia renal*
oil prices	*estructura impositiva*	*frecuencia cardíaca*
domestic demand	*empresa rival*	*alergia respiratoria*
policy mix	*dispersión inicial*	efecto broncodilatador
labour cost	*deseabilidad social*	estado estacionario
job creation	*shock tecnológico*	*tratamiento farmacológico*
interest rates	*impuesto específico*	*poliposis nasal*
support services	*expansión monetaria*	prueba t
price stability	revelación directa	*tracto respiratorio*
ageing populations	*gasto público*	*reactividad bronquial*

Taula 6.8: Distribució de termes per corpus i mesures (II).

Informació mútua		
Econ. ang. JRC	Econ. esp. JRC	Med. esp. IULA
euro area	*Estados miembros*	*asma bronquial*
Lisbon strategy	*mercado interior*	*sexo masculino*
internal market	programas nacionales	*función pulmonar*
human capital	*reformas estructurales*	*sexo femenino*
framework conditions	*contratación pública*	*hiperreactivida bronquial*
public procurement	*políticas macroeconómicas*	grupo control
social partners	*medio ambiente*	consulta externa
productivity growth	*recuperación económica*	capacidad vital
sustainable development	*políticas económicas*	*diabetes mellitus*
industrial base	*capital humano*	*flujo espiratorio*
competitive advantages	*desarrollo sostenible*	*sangre periférica*
structural reforms	*gasto público*	*hipertensión arterial*
energy efficiency	condiciones marco	*volumen espiratorio*
economic recovery	*demanda interna*	*actividad fagocítica*
macroeconomic policies	*finanzas públicas*	pulmonar obstructiva
labour market	*cohesión social*	*insuficiencia renal*
oil prices	sector público	*frecuencia cardíaca*
domestic demand	*productividad laboral*	*alergia respiratoria*
labour cost	sistema económico	efecto broncodilatador
policy mix	*estabilidad macroeconómica*	estado estacionario
job creation	objetivos horizontales	*tratamiento farmacológico*
interest rates	*mercados financieros*	*poliposis nasal*
support services	*crecimiento potencial*	prueba t
price stability	*marco normativo*	*reactividad bronquial*
ageing populations	utilización sostenible	*tracto respiratorio*

Taula 6.9: Distribució de termes per corpus i mesures (III).

***t* de Student**		
Econ. esp. IULA	Econ. ang. JRC	Econ. fr. JRC
política económica	*euro area*	*zone euro*
equilibrio general	*internal market*	*politiques macroéconomiques*
precio neto	Lisbon strategy	*capital humain*
entorno económico	*productivity growth*	base industrielle
elección social	*labour market*	*reprise économique*
sector oligopolístico	*macroeconomic policies*	*demande intérieure*
equilibrio parcial	*employment policies*	*partenaires sociaux*
sector público	public procurement	environnement favorable
problema económico	*social partners*	*politiques économiques*
mecanismo competitivo	sustainable development	*cohésion sociale*
reforma fiscal	*structural reforms*	*protection sociale*
coste marginal	*economic recovery*	*finances publiques*
estructura impositiva	*human capital*	cadre multilatéral
situación inicial	framework conditions	*commerce mondial*
política monetaria	potential growth	environnement concurrentiel
sector exterior	European economy	*stabilité macroéconomique*
informacionalmente eficiente	*employment rate*	coûts environnementaux
competencia perfecta	*industrial base*	coûts administratifs
shock tecnológico	*energy efficiency*	secteurs public
impuesto específico	competitive advantages	*économie mondiale*
dispersión inicial	*oil prices*	incitation économiques
deseabilidad social	support services	conséquences économiques
retorno neto	*price stability*	mesures visant
tipo impositivo	labour cost	*activité économique*
precio bruto	*macroeconomic stability*	croissance potentiel

A més de la distribució dels termes en cada una de les tres franges percentuals analitzades, també convé observar en quin tipus de corpus les mesures situen els termes en els rangs inicials dels resultats, tenint en compte que els corpus d'economia en espanyol, anglès i francès de JRC són els de menor volum, els corpus de serveis socials del Termcat tenen un volum mitjà i els corpus d'economia i medicina en espanyol de l'IULA són els de major volum. En aquest sentit, les mesures que aconsegueixen de situar els termes en els rangs més alts (taula 6.6), distribueixen els termes per corpus de la manera següent:

- El coeficient Dice situa els termes en els rangs inicials d'un dels corpus de menor volum (economia espanyol JRC).
- La ràtio log-likelihood endreça els termes en els rangs inicials de tots tres tipus de corpus: de menor volum (economia anglès JRC), de volum mitjà (serveis socials espanyol Termcat) i de major volum (medicina espanyol IULA).
- La mesura Poisson-Stirling situa els termes en el rang inicial d'un dels corpus de menor volum (economia anglès JRC).
- La mesura informació mútua endreça els termes en els rangs inicials de tots tres tipus de corpus: de menor volum (economia anglès JRC), de volum mitjà (serveis socials espanyol Termcat) i de major volum (medicina espanyol IULA).
- La mesura *t* de Student posa els termes en els rangs inicials de tots tres tipus de corpus: en corpus de menor volum (economia anglès JRC), de volum mitjà (serveis socials espanyol Termcat) i de major volum (economia espanyol IULA).
- El coeficient Jaccard situa els termes en els rangs inicials d'un dels corpus de menor volum (economia francès JRC).

Les dades obtingudes confirmen que les mesures log-likelihood, informació mútua i *t* de Student destaquen per situar els termes en els rangs

inicials de corpus que pertanyen a àmbits temàtics diferents (economia, serveis socials i medicina), els quals també són diversos en volum. Tenint en compte que plantegem l'ús de mesures d'associació lèxica per a millorar la tasca de validació manual de candidats a terme, les mesures que permeten situar un major nombre de termes en els rangs més alts i en un nombre variat de corpus són log-likelihood, informació mútua i *t* de Student. Les mesures Dice, Poisson-Stirling i Jaccard ho aconsegueixen en un sol corpus. En definitiva, constatem que la implementació de la ràtio log-likelihood, la mesura informació mútua i la mesura *t* de Student en el procés d'extracció automàtica de termes facilita la identificació manual de termes d'una llista de candidats per tres motius: primer, perquè situen els termes en els rangs més alts; segon, perquè redueixen el nombre de candidats que ha de ser revisat, i això és degut a l'alta concentració de termes que situen en els rangs inicials dels resultats, i tercer perquè els candidats que situen en les primeres posicions dels resultats corresponen als que tenen una major força de cohesió i, doncs, són els que tenen major probabilitat de ser termes. Els resultats analitzats mostren que és una aportació rellevant per al procés d'extracció automàtica de terminologia la incorporació de les mesures log-likelihood, informació mútua i *t* de Student com a complement a la mesura de freqüència i al procés de validació manual de candidats a terme.

6.2 Avaluació

L'avaluació de les mesures d'associació lèxica aplicades al procés d'extracció automàtica de terminologia se centra a observar la capacitat que tenen aquestes mesures a l'hora d'identificar unitats terminològiques i la rendibilitat que poden oferir en la tasca de validació manual de candidats a terme. Per aquest motiu, l'avaluació de la rendibilitat de les mesures d'associació lèxica aplicada a la tasca de validació manual de candidats extrets amb el mètode de freqüència i el mètode TSR és presentada en termes de precisió i cobertura. Concretament, la precisió mesura l'adequació de les unitats lèxiques proposades com a termes en cada un dels rangs, que correspon a la ràtio entre el nombre de termes correctes (*veritables positius*) i el nombre total d'unitats proposades (*veritables positius* i *falsos positius*). I la cobertura assenyala el nivell amb què els termes són identificats en els corpus, que correspon a la ràtio entre el nombre de termes identificats correctament (*veritables positius*) i el nombre total de termes (*veritables positius* i *falsos negatius*). Per a calcular la precisió i la cobertura que ofereixen les mesures d'associació lèxica, els resultats són contrastats amb el nombre de termes presents en els corpus especialitzats.

Seguidament descrivim l'avaluació de les mesures d'associació lèxica implementades en els resultats obtinguts amb el mètode freqüència i el mètode TSR.

6.2.1 Avaluació de mesures amb el mètode freqüència

Les mesures d'associació lèxica implementades amb el mètode freqüència han estat avaluades tenint en compte el valor de rang atorgat als candidats a terme extrets extrets automàticament dels corpus especialitzats i endreçats per freqüència. L'atribució d'un valor de rang a cada candidat a terme permet disposar d'un resultat d'extracció que es pot basar en la importància d'associació, la força d'associació, la mesura d'informació o l'heurística, segons la mesura d'associació lèxica implementada.

El nombre de termes identificat per cada mesura ha estat contrastat amb la totalitat dels termes presents en els corpus especialitzats. En la taula 6.10 s'especifica el nombre de candidats extrets per corpus i el nombre de termes que hi són presents.

Taula 6.10: Nombre de candidats i termes extrets dels corpus.

Corpus especialitzats	**Llengua corpus**	**Procedència corpus**	**Candidats bigrams**	**Termes bigrams**
Economia	espanyol	JRC (UE)	845	136
Economia	anglès	JRC (UE)	1.881	141
Economia	francès	JRC (UE)	945	120
Economia	espanyol	IULA (UPF)	2.165	517
Serveis socials	català	Termcat	1.661	129
Serveis socials	espanyol	Termcat	1.340	119
Medicina	espanyol	IULA (UPF)	6.740	664

El càlcul de precisió (P) i cobertura (C) obtingut per cada mesura és recollit en les taules 6.11, 6.12, 6.13, 6.14. De cada corpus analitzem unes determinades posicions, de les quals destaquem el resultat més favorable per a la validació manual final dels candidats.

Taula 6.11: Avaluació dels resultats amb el mètode freqüència (I).

Corpus economia espanyol JRC											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
100	P	9	31	20	9	1	32	31	**37**	9	9
	C	6,6	22,7	14,7	6,6	0,7	23,5	22,7	**27,2**	6,6	6,6
500	P	11,8	14,4	11	11	8,4	14,6	14,4	**15,4**	11	43,3
	C	43,3	52,9	40,4	40,4	30,8	53,6	52,9	**56,6**	40,4	18,5
800	P	12,7	12,8	12,5	12,8	11,8	12,8	12,8	**13**	12,8	12,7
	C	75	75,7	73,5	75,7	69,8	75,7	75,7	**76,4**	75,7	75

Taula 6.12: Avaluació dels resultats amb el mètode freqüència (II).

Corpus economia espanyol IULA											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
100	P	7	42	12	7	5	43	42	**47**	7	7
	C	1,3	8,1	2,3	1,3	0,9	8,3	8,1	**9**	1,3	7
500	P	6,8	18	5,2	6,8	4,4	18,2	18,2	**23,4**	6,8	6,8
	C	6,5	17,4	5	6,5	4,2	17,6	17,6	**22,6**	6,5	6,5
1000	P	8,7	12	7,5	8,4	5,3	12,1	12	**13,5**	8,4	8,7
	C	16,8	23,2	14,5	16,2	10,2	23,4	23,2	**26,1**	16,2	16,8
2000	P	10	10,5	10,1	10,3	10	10,4	10,5	**10,6**	10,3	10
	C	38,8	40,6	39,2	40	38,6	40,4	40,6	**41**	40	38,8
Corpus economia anglès JRC											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
100	P	2	26	13	2	0	27	26	**34**	2	2
	C	1,4	18,4	9,2	1,4	0	19,1	18,4	**24,1**	1,4	1,4
500	P	4,8	10,6	3,8	4,6	2,6	10,6	10,8	**14,4**	4,6	4,8
	C	17	37,5	13,4	16,3	9,2	37,5	38,3	**51**	16,3	17
1000	P	4,8	7	4,3	5,2	3,7	7,3	7	**8,4**	5,2	4,8
	C	34	49,6	30,5	36,8	26,2	51,7	49,6	**59,5**	36,8	34
1800	P	6,6	6,6	6,3	6,6	6,3	6,6	6,6	**6,7**	6,6	6,6
	C	84,4	85,1	80,8	84,4	80,8	84,4	85,1	**85,8**	84,4	84,4
Corpus economia francès JRC											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
100	P	9	24	16	9	3	24	24	**25**	9	9
	C	7,5	20	13,3	7,5	2,5	20	20	**20,8**	7,5	7,5
500	P	8	9,4	7,4	7,6	6	9,6	9,4	**10,2**	7,6	8
	C	33,3	39,1	30,8	31,6	25	40	39,1	**42,5**	31,6	33,3
800	P	8,6	8,8	8,5	8,6	7,7	9	8,6	**9,1**	8,6	8,6
	C	57,5	59,1	56,6	57,5	51,6	60	57,5	**60,8**	57,5	57,5

Taula 6.13: Avaluació dels resultats amb el mètode freqüència (III).

Corpus serveis socials català Termcat											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
100	P	1	11	5	1	0	11	11	**15**	1	1
	C	0,7	8,5	3,8	0,7	0	8,5	8,5	**11,6**	0,7	0,7
500	P	3	5,6	2,8	3,2	1,6	5,6	5,6	**6**	3,2	3
	C	11,6	21,7	10,8	12,4	6,2	21,7	21,7	**23,2**	12,4	11,6
800	P	2,7	**4,5**	2,5	3	2,1	4,3	**4,5**	4,2	3	2,7
	C	17	**27,9**	15,5	18,6	13,1	27,1	**27,9**	26,3	18,6	17
1000	P	2,9	**3,8**	2,9	3	2,2	3,7	**3,8**	3,7	3	2,9
	C	22,4	**29,4**	22,4	23,2	17	28,6	**29,4**	28,6	23,2	22,4
Corpus serveis socials espanyol Termcat											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
100	P	1	14	6	1	1	14	14	**19**	1	1
	C	0,8	11,7	5	0,8	0,8	11,7	11,7	**15,9**	0,8	0,8
500	P	4	6,6	3,8	3,4	1,8	6,4	6,6	**7,2**	3,4	4
	C	16,8	27,7	15,9	14,2	7,5	26,8	27,7	**30,2**	14,2	16,8
800	P	3,3	4,8	3,2	3,5	3,3	4,6	4,7	**5,1**	3,5	3,3
	C	22,6	32,7	21,8	23,5	22,6	31	31,9	**34,4**	23,5	22,6
1000	P	3,5	**4,7**	3,2	3,7	3	4,6	**4,7**	**4,7**	3,7	3,5
	C	29,4	**39,5**	26,8	31	25,2	38,6	**39,5**	**39,5**	31	29,4

Taula 6.14: Avaluació dels resultats amb el mètode freqüència (IV).

Corpus medicina espanyol IULA										
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
100	P	7	17	8	7	5	18	17	**21**	7	7
	C	1	2,5	1,2	1	0,7	2,7	2,5	**3,1**	1	1
500	P	4,2	13,4	5,8	4,2	3,8	12,4	13,4	**14**	4,2	4,2
	C	3,1	10	4,3	3,1	2,8	9,3	10	**10,5**	3,1	3,1
1000	P	4,9	9,5	4,7	4,9	3,8	9,7	9,4	**10,6**	4,7	4,9
	C	7,3	14,3	7	7,3	5,7	14,6	14,1	**15,9**	7	7,3
3000	P	4,4	5,9	4,1	4,4	3,6	5,8	5,7	**6,1**	4,4	4,4
	C	20	26,6	18,8	20,1	16,5	26,3	26	**27,7**	20	20
5000	P	4,2	4,6	4	4,3	3,8	4,6	**4,7**	**4,7**	4,3	4,2
	C	31,6	35	30,4	32,8	28,7	35	**35,5**	**35,8**	32,8	31,6

En les taules 6.11, 6.12, 6.13, 6.14 hem calculat el nivell de precisió i cobertura que assoleixen les mesures d'associació lèxica amb relació al posicionament dels termes en una llista de candidats. Així, per exemple, del corpus d'economia espanyol (JRC) hem avaluat les posicions 100, 500 i 800 i hem obtingut uns resultats de precisió i cobertura de la mesura *t* de Student destacats respecte de la resta de mesures.

Els resultats obtinguts en tots els corpus confirmen que la mesura *t* de Student permet endreçar els candidats a terme en una posició superior en rang en comparació amb la resta de mesures d'associació lèxica analitzades. La precisió i la cobertura obtingudes amb aquesta mesura es manté estable i en posició preeminent en tots els corpus especialitzats, els quals són d'àmbit, llengua i volum diferent.

6.2.2 Avaluació de mesures amb el mètode TSR

L'avaluació dels resultats obtinguts amb les mesures d'associació lèxica implementades en el procés d'extracció automàtica de terminologia basat en el mètode TSR l'hem centrada en el càlcul de precisió i cobertura assolit per les mesures amb relació al nombre de termes presents en els corpus de prova.

La distribució del nombre de candidats i termes que han estat extrets dels corpus de prova queda recollida en la taula 6.15.

Taula 6.15: Distribució de candidats i termes per corpus.

Corpus especialitzats	**Llengua corpus**	**Procedència corpus**	**Candidats prova**	**Termes prova**
Economia	espanyol	JRC (UE)	157	80
Economia	anglès	JRC (UE)	365	97
Economia	francès	JRC (UE)	96	76
Economia	espanyol	IULA (UPF)	414	211
Serveis socials	català	Termcat	80	62
Serveis socials	espanyol	Termcat	86	61
Medicina	espanyol	IULA (UPF)	587	307

El càlcul de precisió i cobertura l'hem determinat a partir de les posicions que ocupen els candidats a terme. En les taules 6.16, 6.17, 6.18, 6.19 recollim les diferents posicions analitzades i els corresponents valors de precisió i cobertura classificats per corpus i mesures. Així mateix, de cada posició destaquem la mesura que ha obtingut més bons resultats.

Taula 6.16: Avaluació dels resultats amb el mètode TSR (I).

Corpus economia espanyol JRC (UE)											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
10	P	50	**90**	80	50	20	**90**	**90**	**90**	50	50
	C	6,2	**11,2**	10	6,2	2,5	**11,2**	**11,2**	**11,2**	6,2	6,2
50	P	48	52	46	48	40	52	52	**56**	48	48
	C	30	32,5	28,7	30	25	32,5	32,5	**35**	30	30
100	P	**42**	41	39	41	37	41	41	**42**	41	**42**
	C	**52,5**	51,2	48,7	51,2	46,2	51,2	51,2	**52,5**	51,2	**52,5**
150	P	**36,6**	**36,6**	36	**36,6**	36	**36,6**	**36,6**	**36,6**	**36,6**	**36,6**
	C	**68,7**	**68,7**	67,5	**68,7**	67,5	**68,7**	**68,7**	**68,7**	**68,7**	**68,7**
Corpus economia anglès JRC (UE)											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
10	P	**70**	60	60	**70**	40	60	60	**70**	**70**	**70**
	C	**7,2**	6,1	6,1	**7,2**	4,1	6,1	6,1	**7,2**	**7,2**	**7,2**
50	P	36	50	30	34	20	50	50	**54**	34	36
	C	18,5	25,7	15,4	17,5	10,3	25,7	25,7	**27,8**	17,5	18,5
100	P	32	35	26	27	20	36	35	**40**	27	32
	C	32,9	36	26,8	27,8	20,6	37,1	36	**41,2**	27,8	32,9
200	P	26	**28**	23,5	25,5	22	26,5	**28**	27,5	25,5	26
	C	53,6	**57,7**	48,4	52,5	45,3	54,6	**57,7**	56,7	52,5	53,6
350	P	22	22	22	22	22	**22,2**	22	22	22	22
	C	79,3	79,3	79,3	79,3	79,3	**80,4**	79,3	79,3	79,3	79,3

Taula 6.17: Avaluació dels resultats amb el mètode TSR (II).

Corpus economia francès JRC (UE)											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
10	P	60	70	70	60	60	70	70	**80**	60	60
	C	7,8	9,2	9,2	7,8	7,8	9,2	9,2	**10,5**	7,8	7,8
50	P	**50**	48	46	48	44	48	48	48	48	**50**
	C	**32,8**	31,5	30,2	31,5	28,9	31,5	31,5	31,5	31,5	**32,8**
90	P	**43,3**	42,2	**43,3**	42,2	**43,3**	42,2	42,2	**43,3**	42,2	**43,3**
	C	**51,3**	50	**51,3**	50	**51,3**	50	50	**51,3**	50	**51,3**
Corpus economia espanyol IULA (UPF)											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
10	P	60	90	50	50	40	90	90	**100**	50	60
	C	2,8	4,2	2,3	2,3	1,9	4,2	4,2	**4,7**	2,3	2,8
50	P	52	80	40	46	26	78	80	**84**	46	52
	C	12,3	18,9	9,4	10,9	6,1	18,4	18,9	**19,9**	10,9	12,3
100	P	48	59	43	48	26	58	59	**68**	48	48
	C	22,7	27,9	20,3	22,7	12,3	27,4	27,9	**32,2**	22,7	22,7
150	P	41,3	47,3	36,6	43,3	32	48	47,3	**54**	43,3	41,3
	C	29,3	33,6	26	30,8	22,7	34,1	33,6	**38,3**	30,8	29,3
200	P	38	45	38	40	31	45,5	45	**48**	40	38
	C	36	42,6	36	37,9	29,3	43,1	42,6	**45,5**	37,9	36
300	P	37,3	**38,6**	36,6	38,3	35,3	38,3	**38,6**	38	38,3	37,3
	C	53	**54,9**	52,1	54,5	50,2	54,5	**54,9**	54	54,5	53
400	P	**37,2**	**37,2**	37	37	37	37	37	**37,2**	37	**37,2**
	C	**70,6**	**70,6**	70,1	70,1	70,1	70,1	70,1	**70,6**	70,1	**70,6**

Taula 6.18: Avaluació dels resultats amb el mètode TSR (III).

Corpus serveis socials català Termcat											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
10	P	30	60	30	30	0	**70**	60	60	30	30
	C	4,8	9,6	4,8	4,8	0	**11,2**	9,6	9,6	4,8	4,8
50	P	28	**30**	28	28	26	28	**30**	**30**	28	28
	C	22,5	**24,1**	22,5	22,5	20,9	22,5	**24,1**	**24,1**	22,5	22,5
75	P	33,3	32	**34,6**	32	**34,6**	**34,6**	32	**34,6**	32	33,3
	C	40,3	38,7	**41,9**	38,7	**41,9**	**41,9**	38,7	**41,9**	38,7	40,3
Corpus serveis socials espanyol Termcat											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
10	P	40	**60**	40	40	10	50	**60**	**60**	40	40
	C	6,5	**9,8**	6,5	6,5	1,6	8,2	**9,8**	**9,8**	6,5	6,5
50	P	34	**42**	36	38	36	36	**42**	40	38	34
	C	27,8	**34,4**	29,5	31,1	29,5	29,5	**34,4**	32,7	31,1	27,8
80	P	40	41,2	**42,5**	41,2	**42,5**	40	41,2	**42,5**	41,2	40
	C	52,4	54,1	**55,7**	54,1	**55,7**	52,4	54,1	**55,7**	54,1	52,4

Taula 6.19: Avaluació dels resultats amb el mètode TSR (IV).

Corpus medicina espanyol IULA											
Top	P/C	Dice	Ll	Odds	Phi	Pmi	Ps	Im	*t* St.	X^2	Jacc.
10	P	30	60	20	40	30	60	60	**80**	40	30
	C	0,9	1,9	0,6	1,3	0,9	1,9	1,9	**2,6**	1,3	0,9
50	P	36	48	32	34	30	46	48	**52**	34	36
	C	5,8	7,8	5,2	5,5	4,8	7,4	7,8	**8,4**	5,5	5,8
100	P	35	**42**	33	33	22	39	**42**	39	33	35
	C	11,4	**13,6**	10,7	10,7	7,1	12,7	**13,6**	12,7	10,7	11,4
200	P	28	30,5	23,5	30,5	22,5	31	30,5	**33,5**	30,5	28
	C	18,2	19,8	15,3	19,8	14,6	20,2	19,8	**21,8**	19,8	18,2
300	P	24,6	28,6	22,6	23,3	20,3	26,6	28,3	**29,6**	23,3	24,6
	C	24,1	28	22,1	22,8	19,8	26	27,6	**28,9**	22,8	24,1
400	P	23,5	24	21,7	22,5	20,7	23,5	24,2	**25**	22,5	23,2
	C	30,2	31,2	28,3	29,3	27	30,6	31,6	**32,5**	29,3	30,2
600	P	21,8	22,3	20,8	**22,5**	20,3	22,1	22,3	22	**22,5**	21,8
	C	42,6	43,6	40,7	**43,9**	39,7	43,3	43,6	43	**43,9**	42,6
750	P	20,1	20,9	20,6	20,9	20,6	20,6	**21**	20,8	20,9	20,1
	C	49,1	51,1	50,4	51,1	50,4	50,4	**51,4**	50,8	51,1	49,1

En les taules 6.16, 6.17, 6.18, 6.19 hem calculat el nivell de precisió i cobertura que assoleixen les mesures d'associació lèxica en ser implementades en el procés d'extracció basat en el mètode TSR. Els resultats obtinguts mostren la rendibilitat que assoleixen les mesures amb relació al posicionament dels termes en una llista de candidats. En aquest sentit observem que en el corpus d'economia espanyol JRC en la posició 10 assoleixen un major nivell de precisió i cobertura les mesures log-likelihood, Poisson-Stirling, informació mútua i *t* de Student; en la posició 50 solament la mesura *t* de Student; en la posició 100, les mesures dice, *t* de

Student i Jaccard, i en la posició 150 totes les mesures, excepte odds i pointwise mutual information, assoleixen un major valor de precisió i cobertura. En conjunt, la mesura *t* de Student és l'única que assoleix una major precisió i cobertura en totes les posicions analitzades. Si observem el corpus d'economia anglès JRC, la mesura *t* de Student també assoleix un major nivell de precisió i cobertura en tres de les cinc posicions analitzades, i en les dues posicions restants els resultats obtinguts són pròxims als millors resultats obtinguts per altres mesures. En el corpus d'economia francès JRC les mesures dice, *t* de Student i Jaccard són les que assoleixen resultats destacats en termes de precisió i cobertura. Pel que fa al corpus d'economia espanyol IULA la mesura *t* de Student destaca per obtenir el nivell més alt de precisió i cobertura en totes les posicions. En el corpus de serveis socials català Termcat les mesures Poisson-Stirling i *t* de Student obtenen resultats destacats en dues de les tres posicions analitzades, i en el corpus de serveis socials espanyol Termcat destaquen les mesures log-likelihood, informació mútua i *t* de Student. I, finalment, en el corpus de medicina espanyol IULA la mesura *t* de Student obté un resultat de precisió i cobertura superior al de la resta de mesures en sis de les vuit posicions, i en les dues posicions restants el nivell de resultats s'acosta molt als valors més alts. En definitiva, constatem que hi ha un grup de sis mesures que obtenen resultats destacats en termes de precisió i cobertura i que són dice, log-likelihood, informació mútua, *t* de Student i Jaccard. Ara bé, d'aquestes mesures únicament *t* de Student obté un millor percentatge de precisió i cobertura en tots set corpus analitzats, ja que les cinc mesures restants assoleixen resultats destacats en un sol corpus.

Els resultats que han obtingut les mesures d'associació lèxica indiquen que no han estat influïts per la mida o l'àmbit d'especialitat dels corpus, fet que queda reflectit en una cobertura i precisió uniformes en tots els corpus. En aquest sentit, el fet de poder contrastar resultats provinents d'àmbits d'especialitat variats i de mides diferents, ens permet constatar la no influència de la mida del corpus ni el tipus de contingut en els resultats assolits per cada una de les mesures.

Des del punt de vista de la rendibilitat en l'aplicació de mesures d'associació lèxica en la tasca de validació manual de candidats a terme, els resultats obtinguts indiquen una significativa concentració en nombre de termes en els rangs inicials, especialment destacada en el cas de la mesura *t* de Student. Concretament observem que aquesta mesura és la que concentra més termes en la majoria de posicions del conjunt de corpus especialitzats que han estat analitzats. En conseqüència, constatem que *t* de Student és la mesura més rendible per a ser incorporada en un procés d'extracció automàtica de terminologia basat en el mètode TSR per tal de poder facilitar el procés de validació manual final de candidats a terme.
Completem l'anàlisi quantitativa dels resultats obtinguts amb les mesures d'associació lèxica combinades amb el mètode TSR i el mètode freqüència fent una anàlisi qualitativa dels candidats extrets en les posicions inicials en els set corpus especialitzats. Concretament, en les taules 6.20, 6.21, 6.22, 6.23 mostrem quins són els candidats i els termes que hem obtingut en les posicions inicials a partir de l'extracció automàtica feta amb ambdós mètodes combinats amb *t* de Student, la mesura que obté més bons resultats en la distribució dels termes per rangs en les llistes de candidats. Fem un recull dels primers candidats a terme extrets de la totalitat de corpus analitzats en el nostre estudi per tal de mostrar com queden distribuïts els termes per rangs en cada un dels corpus.

La tendència que es confirma en el conjunt dels corpus és un major endreçament de termes en les diferents posicions dels resultats quan el mètode TSR es combina amb la mesura *t* de Student, més que no pas quan el mètode freqüència es combina amb aquesta mesura. A partir d'aquí podem concretar que en els corpus d'economia francès de JRC, medicina espanyol de l'IULA, serveis socials espanyol del Termcat i economia anglès de JRC s'observa que els termes estan situats en un rang superior i, en conseqüència, la diferència en la identificació de termes entre el mètode TSR i el mètode freqüència és major.

Així mateix, malgrat constatar en l'anàlisi quantitativa que els dos mètodes combinats amb la mesura *t* de Student obtenen uns resultats rellevants en l'avaluació de les dades analitzades, en la mostra d'exemples que aportem s'observa que el mètode TSR redueix força l'aparició de candidats que contenen combinacions de paraules genèriques que no aporten cap tipus de valor als resultats, i que són candidats que acostumen a ser força presents en els resultats que s'obtenen amb el mètode freqüència. Aquest fet es deu al filtratge de candidats que fa el mètode TSR a partir de *tokens* terminològics. Així, en el filtratge de freqüència del corpus d'economia espanyol JRC trobem *Directriz n*, *medio plazo*, *largo plazo*; en el corpus d'economia anglès JRC hi ha *Lisbon strategy*, *same time*, *integrated guideline*, *medium term*; en el corpus d'economia francès JRC observem *bon fonctionnement*, *grandes orientations*, *faible niveau*, *second semestre*; en el corpus d'economia espanyol IULA hi ha *generaciones sucesivas*; en el corpus de serveis socials en català i espanyol observem *presente ley*, *departamento competente*, *mil habitantes*, *socials bàsics*, *departament competent*, i en el corpus de medicina IULA hi ha *grupo control*, *doble ciego*, *presente estudio*. I en el filtratge de TSR observem que també hi ha algun tipus d'aquests candidats, com ara *environnement favorable*, *informacionalmente eficiente*, *area basica* o *grupo II*. La capacitat que té la mesura *t* de Student de situar en el nivell de rang més alt els termes presents en un corpus especialitzat, quan és combinada amb els candidats extrets amb el mètode TSR, permet millorar els reendreçament dels termes i situar-los en les posicions inicials dels resultats, i, per tant, la revisió manual final dels candidats queda focalitzada en els resultats més rellevants i hi ha un estalvi en la revisió de candidats buits de contingut.

Convé afegir que la combinació del mètode TSR amb la mesura *t* de Student permet disposar de l'ordre de rang dels candidats que tenen major valor terminològic, ja que són els candidats que apareixen en les posicions destacades dels resultats pel fet d'estar constituïts per *tokens* que pertanyen a l'àrea d'especialitat del corpus i també perquè l'ordre de rang de la mesura indica que aquests *tokens* es caracteritzen per la seva cohesió lèxica. Així, doncs, s'incorpora el valor de la pertinença terminològica

dels candidats al corpus especialitzat a l'hora d'obtenir l'ordre final dels candidats, un valor determinant per a la posterior identificació manual dels termes.

La mostra qualitativa de resultats que aportem a continuació confirma el comportament similar que té el mètode TSR combinat amb la mesura *t* de Student en tots els corpus especialitzats respecte l'endreçament dels termes. En aquest sentit, doncs, podem confirmar la hipòtesi de partida en la qual volíem constatar el nivell de rendibilitat de les mesures d'associació lèxica a l'hora de situar els termes en les posicions inicials d'una llista de candidats i, en conseqüència, reduir el nombre de candidats que ha de ser validat manualment al final del procés d'extracció de termes. En l'estudi que hem dut a terme hem constatat la rellevància de la mesura *t* de Student, la qual, combinada amb el mètode TSR, permet situar per ordre de rang un major nombre de termes en les posicions inicials d'una llista de candidats. Així mateix, en el nostre estudi hem observat que les mesures log-likelihood i informació mútua també permeten reduir significativament el nombre de candidats que ha de ser revisat manualment en la totalitat de la llista de resultats. En conseqüència, doncs, podem confirmar que el mètode TSR és el més adequat per a ser combinat en primera instància amb la mesura *t* de Student, i també amb les mesures log-likelihood i informació mútua, a fi de reduir en temps i cost la tasca que ha de dur a terme un especialista en la revisió manual de candidats.

Taula 6.20: Comparació de TSR i freqüència amb *t* de Student (I).

TSR amb *t* de Student		Freqüència amb *t* de Student	
Corpus economia espanyol JRC			
Estados miembros	T	Estados miembros	T
mercado interior	T	políticas macroeconómicas	T
políticas macroeconómicas	T	Directriz n	CT
programas nacionales	CT	políticas económicas	T
políticas económicas	T	reformas estructurales	T
medio ambiente	T	Consejo Europeo	CT
reformas estructurales	T	mercado interior	T
contratación pública	T	medio plazo	CT
recuperación económica	T	programas nacionales	CT
desarrollo sostenible	T	largo plazo	CT
crecimiento económico	T	medio ambiente	T
crecimiento potencial	T	recuperación económica	T
capital humano	T	crecimiento económico	T
gasto público	T	interlocutores sociales	CT
condiciones marco	CT	contratación pública	T
sector público	CT	gasto público	T
cohesión social	T	desarrollo sostenible	T
Corpus economia anglès JRC			
euro area	T	Member States	CT
internal market	T	euro area	T
Lisbon strategy	CT	growth potential	CT
productivity growth	T	Lisbon strategy	CT
labour market	T	European Council	CT
macroeconomic policies	T	internal market	T
employment policies	T	labour market	T
public procurement	CT	productivity growth	T
social partners	CT	macroeconomic policies	T
sustainable development	CT	same time	CT
structural reforms	T	integrated guideline	CT
economic recovery	T	structural reforms	T
human capital	T	employment rate	T
framework conditions	CT	medium term	CT
potential growth	CT	employment policies	T
European economy	CT	economic recovery	T
employment rate	T	efficient allocation	CT

Taula 6.21: Comparació de TSR i freqüència amb *t* de Student (II).

TSR amb *t* de Student		Freqüència amb *t* de Student	
Corpus economia francès JRC			
zone euro	T	États membres	CT
politiques macroéconomiques	T	lignes directrices	CT
capital humain	T	croissance économique	T
base industrielle	CT	zone euro	T
reprise économique	T	politiques macroéconomiques	T
demande intérieure	T	Union européenne	CT
partenaires sociaux	T	niveau national	CT
environnement favorable	CT	politiques économiques	T
politiques économiques	T	marché intérieur	T
cohésion sociale	T	programmes nationaux	CT
protection sociale	T	ligne directrice	CT
finances publiques	T	bon fonctionnement	CT
cadre multilatéral	CT	infrastructures européennes	CT
commerce mondial	T	grandes orientations	CT
environnement concurrentiel	CT	long terme	CT
stabilité macroéconomique	T	base industrielle	CT
coûts environnementaux	CT	faible niveau	CT
Corpus economia espanyol IULA			
política económica	T	coste marginal	T
equilibrio general	T	política económica	T
precio neto	T	equilibrio general	T
entorno económico	T	dinero fiduciario	T
elección social	T	generaciones sucesivas	CT
sector oligopolístico	T	precio neto	T
equilibrio parcial	T	entorno económico	T
sector público	T	agentes económicos	T
problema económico	T	elección social	T
mecanismo competitivo	T	sector oligopolístico	T
reforma fiscal	T	ingreso marginal	T
coste marginal	T	equilibrio parcial	T
estructura impositiva	T	competencia perfecta	T
situación inicial	CT	sectores productivos	CT
política monetaria	T	fluctuaciones económicas	CT
sector exterior	T	entornos económicos	T
informacionalmente eficiente	CT	ciclos económicos	CT

Taula 6.22: Comparació de TSR i freqüència amb *t* de Student (III).

TSR amb *t* de Student		Freqüència amb *t* de Student	
Corpus serveis socials espanyol Termcat			
discapacidad intelectual	T	servicios sociales	T
violencia machista	T	presente ley	CT
iniciativa social	T	entes locales	CT
financiación pública	CT	Consejo General	CT
asistente personal	T	sistema público	CT
innovación tecnológica	CT	departamento competente	CT
titularidad pública	T	iniciativa social	T
atención social	T	atención social	T
iniciativa mercantil	CT	Atención Pública	CT
inserción SOI	CT	personal profesional	CT
acceso universal	T	discapacidad intelectual	T
estabilidad laboral	CT	Plan estratégico	CT
protección provisional	CT	personas mayores	T
teleasistencia domiciliaria	T	administraciones públicas	CT
protección jurídica	T	mil habitantes	CT
centro residencial	T	Información Social	CT
asistencia tecnológica	T	prestaciones económicas	T
Corpus serveis socials català Termcat			
discapacitat física	T	serveis socials	T
gent gran	T	socials bàsics	CT
violència masclista	T	Consell General	CT
problemàtica social	CT	sistema públic	CT
mòdul social	T	departament competent	CT
titularitat pública	T	socials especialitzats	CT
participació cívica	T	discapacitat física	T
àrea bàsica	CT	personal professional	CT
teleassistència domiciliària	CT	Pla estratègic	CT
protecció provisional	CT	administracions públiques	CT
centre residencial	T	gent gran	T
protecció jurídica	T	prestacions econòmiques	T
xarxa pública	CT	Sistema Català	CT
iniciativa privada	CT	cohesió social	T
administració pública	CT	malaltia mental	T
risc greu	CT	àrees bàsiques	CT
participació comunitària	CT	normativa reguladora	CT

Taula 6.23: Comparació de TSR i freqüència amb *t* de Student (IV).

Corpus medicina espanyol IULA			
TSR amb *t* de Student		Freqüència amb *t* de Student	
asma bronquial	T	asma bronquial	T
hiperreactividad bronquial	T	hiperreactividad bronquial	T
función pulmonar	T	función pulmonar	T
grupo control	CT	grupo control	CT
sexo masculino	T	efectos colaterales	CT
sexo femenino	T	doble ciego	CT
flujo espiratorio	T	sexo masculino	T
volumen espiratorio	T	obstrucción bronquial	T
obstrucción bronquial	T	flujo espiratorio	T
consulta externa	CT	provocación bronquial	CT
capacidad vital	CT	presente estudio	CT
reactividad bronquial	T	mg ml	CT
provocación bronquial	CT	IgE total	CT
diabetes mellitus	T	efectos secundarios	CT
alergia respiratoria	T	efecto broncodilatador	CT
asma leve	T	antecedentes familiares	CT
enfermedad respiratoria	T	sexo femenino	T
grupo II	CT	mg kg	CT
sangre periférica	T	volumen espiratorio	T
hipertensión arterial	T	reactividad bronquial	T

6.3 Conclusions

En el present capítol hem presentat la capacitat que tenen les mesures d'associació lèxica descrites en el capítol 5 per a identificar els termes que hi ha presents en una llista de candidats extrets d'un corpus especialitzat per tal de fer més rendible la tasca de validació manual dels candidats. Per a fer-ho, hem implementat les mesures en dos processos diferents d'extracció de candidats a terme: l'un basat en el mètode freqüència, en el qual els candidats són extrets a partir de la freqüència d'aparició en el corpus (apartat 6.1.1), i l'altre basat en el mètode TSR, que fa ús de *tokens* terminològics per a extreure candidats a terme d'un corpus especialitzat (apartat 6.1.2).

Si tenim en compte els resultats obtinguts a partir del mètode freqüència constatem que *t* de Student és la mesura que situa un major nombre de termes en els rangs més alts de tots els corpus analitzats. Així mateix, observem que el nombre de termes identificats amb la freqüència d'aparició al corpus i amb les mesures d'associació lèxica és força similar. D'aquests resultats es desprèn que la informació de freqüència i rang obtinguda dels candidats facilita la identificació dels termes, ja que com més alta sigui la freqüència amb què apareix un candidat en un corpus i també el valor de rang que li hagi assignat la mesura *t* de Student més probabilitat tindrà aquesta unitat de ser terminològica.

Amb relació a la implementació de les mesures al mètode TSR, també hem comparat els resultats d'extracció de termes obtingut inicialment amb el mètode TSR i amb les diferents mesures en diferents posicions. Els resultats constaten que *t* de Student continua essent la mesura que identifica un major nombre de termes en les diferents posicions, seguida de les mesures log-likelihood i informació mútua. També observem que els resultats d'extracció obtinguts amb el mètode TSR són similars als de la mesura *t* de Student. En aquest sentit, doncs, per a tots dos mètodes hi ha una mateixa mesura que obté uns resultats destacats.

Hem completat l'anàlisi de resultats obtinguts amb les mesures d'associació lèxica a partir de les posicions en què queden endreçats els candidats amb un estudi de la rendibilitat que assoleixen les mesures pel que fa a la tasca de validació manual de candidats a terme. En aquest sentit, de cada mesura hem analitzat el nombre de candidats que s'ha de revisar manualment per a poder identificar el 25%, el 50% i el 75% dels termes presents en cada corpus. Per a fer-ho, hem identificat el nombre de termes que hi ha en cada corpus i els candidats que se n'han extret, i hem calculat el nombre de candidats que s'ha de revisar per a obtenir els termes de cada franja percentual. Aquest estudi ens ha permès constatar que les mesures log-likelihood, informació mútua i *t* de Student identifiquen el major nombre de termes en cada una de les franges percentuals havent de revisar el menor nombre de candidats (taula 6.6). A més, també hem pogut confirmar que aquestes tres mesures obtenen els millors resultats en corpus de diferent volum i àmbit temàtic (economia, serveis socials i medicina). Així, doncs, aquests resultats constaten un ús efectiu de les mesures log-likelihood, informació mútua i *t* de Student per a millorar la tasca de validació manual de candidats a terme.

Seguidament hem avaluat la proposta experimental proposada tenint en compte els resultats obtinguts amb el mètode de freqüència i el mètode TSR. Per a fer-ho, hem calculat la precisió i la cobertura que ofereixen les diferents mesures d'associació lèxica a partir d'unes determinades posicions en què se situen els candidats. Les dades obtingudes de l'avaluació amb el mètode freqüència mostren que *t* de Student és la mesura que assoleix una major precisió i cobertura en endreçar la major part dels termes en els rangs més alts. Aquests resultats s'obtenen en diferents corpus, d'àmbit, llengua i volum diferents. Les dades obtingudes amb el mètode TSR indiquen que hi ha sis mesures que mostren uns resultats destacats i que són dice, log-likelihood, informació mútua, *t* de Student i Jaccard, de les quals *t* de Student és la mesura que obté el millor resultat de precisió i cobertura en tots els set corpus analitzats. Les cinc mesures restants sols obtenen bons resultats en un dels corpus.

Com a conclusió podem afirmar que *t* de Student, juntament amb log-likelihood i informació mútua, són les mesures més rendibles per a ser implementades en un procés d'extracció automàtica de terminologia amb l'objectiu de facilitar la tasca de validació manual final dels candidats a terme extrets d'un corpus especialitzat. Així mateix, la rendibilitat assolida per aquestes mesures confirma els resultats que s'han obtingut en treballs de recerca previs, en els quals s'utilitzen mètodes estadístics per a identificar automàticament termes presents en corpus de caràcter especialitzat (Daille, 1997; Evert i Krenn, 2005; Pazienza *et al.*, 2005; Boulaknadel *et al.*, 2008). I també consolida els resultats que vam obtenir en el projecte de recerca (Vàzquez i Oliver, 2007).

Capítol 7

CONCLUSIONS

L'interès per la identificació automàtica de les unitats terminològiques que són presents en àmbits especialitzats sorgeix en paral·lel a l'aparició dels primers ordinadors amb capacitat per a processar un volum important de dades. Així va néixer una nova manera de treballar i relacionar-se amb els textos que permet processar corpus textuals de gran volum per a extreure'n les dades més significatives, com ara els termes.

Aquest nou context ha fet possible establir diferents estratègies de processament dels corpus per a arribar a extreure'n les unitats més adequades segons l'objecte d'estudi. Concretament, en l'àmbit de la terminologia s'han desenvolupat mètodes d'extracció centrats en la identificació de termes monoparaula (*monoword terms*) i termes multiparaula (*multiword terms*), expressions formades per una paraula o més d'una que tenen una estructura gramatical i un significat específic (Maynard i Ananiadou, 1999; Frantzi *et al.*, 2000; Maynard i Ananiadou, 2000b; Huo, 2012), que es diferencien dels mètodes que tenen per objecte la selecció d'expressions multiparaula (*multiword expressions*), les quals corresponen a sintagmes, locucions o noms propis presents en un text (Calzolari *et al.*, 2002; Copestake *et al.*, 2002).

Els mètodes d'extracció automàtica de termes que s'implementen en eines de processament de corpus textuals tendeixen a extreure un gran nombre d'unitats que són candidates a ser termes, malgrat els avenços aconseguits en l'anàlisi i detecció d'unitats terminològiques, fet que dificulta la identificació manual dels termes propis d'un àmbit d'especialitat per part d'un especialista. Aquesta limitació és pròpia dels mètodes que apliquen estratègies lingüístiques, estadístiques i híbrides en el procés d'extracció de candidats. Per aquest motiu, en el present treball de recerca hem analitzat una combinació d'estratègies estadístiques amb l'objectiu de comprovar si tenen capacitat per a identificar d'un corpus els candidats amb major probabilitat de ser terminològics, i així reduir en nombre els candidats que han de ser validats manualment al final del procés d'extracció.

Una de les estratègies estadístiques que hem analitzat per a poder identificar els candidats amb més probabilitat de ser termes, es basa en l'aprofitament dels termes propis d'un àmbit d'especialitat amb l'objectiu de poder-ne detectar de nous d'una manera recursiva, i que anomenem *mètode token slot recognition*. Aquest mètode ens ha permès constatar la rendibilitat que ofereix l'extracció de candidats a partir del nivell de coincidència que presenten amb els termes propis d'un àmbit d'especialitat, en comparació de l'extracció no recursiva de candidats basada en la freqüència. Aplicant aquest mètode hem aconseguit identificar un major nombre de termes respecte dels resultats de freqüència, limitar el nombre de candidats extrets a aquelles unitats que comparteixen *tokens* terminològics amb els termes de referència i també poder ampliar el nombre de candidats amb caràcter terminològic per mitjà de l'extracció recursiva de candidats. En aquest cas, la recursivitat del mètode ens permet primer validar un candidat i després, si escau, incorporar-lo com a terme de referència per a poder filtrar nous candidats. Així mateix, convé observar que aquest mètode també es pot implementar encara que no disposem de termes de referència d'un àmbit d'especialitat. En aquest cas, es constitueix la llista de termes de referència a mesura que es validen els candidats extrets. Els resultats que hem obtingut de l'aplicació del mètode *token slot recognition* en diferents corpus especialitzats confirmen la hipòtesi

de partida en la qual indicàvem que *un mètode recursiu d'extracció automàtica de termes basat en estratègies estadístiques permet recuperar un major nombre de termes que un mètode no recursiu basat en la freqüència*.

Una altra estratègia estadística avaluada ha estat la implementació d'onze mesures d'associació lèxica en el procés d'extracció automàtica de termes amb l'objectiu de reduir el nombre de candidats que s'ha de revisar manualment al final del procés d'extracció. En aquest sentit, hem pogut comprovar que quan les mesures són implementades en un procés d'extracció basat en la freqüència, aquestes no aconsegueixen millorar el nombre de termes extrets. De la mateixa manera, hem constatat que quan les mesures són implementades amb el mètode *token slot recognition*, aleshores tenen capacitat per a detectar un major nombre de termes. Així, si ens centrem en la rendibilitat que ofereixen les mesures analitzades amb relació a la seva capacitat de reduir en nombre els candidats que s'han de validar manualment al final del procés d'extracció, podem afirmar que les mesures d'associació lèxica aconsegueixen millors resultats amb el mètode *token slot recognition* i que hi ha tres mesures –ràtio log-likelihood, mesura informació mútua i prova *t* de Student– que situen un major nombre de termes en les posicions inicials dels resultats. Amb aquestes tres mesures s'aconsegueix de reduir el volum de candidats que s'ha de revisar manualment al final d'un procés d'extracció automàtica de termes. Els resultats obtinguts d'avaluar aquesta estratègia relacionada amb la validació manual dels candidats confirmen la hipòtesi plantejada inicialment: *les mesures d'associació lèxica permeten disposar d'una llista de candidats endreçats de major a menor probabilitat de ser termes i, en conseqüència, redueixen el nombre de candidats que ha de ser validat manualment al final d'un procés d'extracció automàtica de termes*. D'aquesta hipòtesi inicial convé matisar que no totes les mesures són igualment rendibles per a la tasca de validació de candidats, ja que en el nostre estudi n'identifiquem tres com a més rendibles, i que aquestes tres mesures aconsegueixen millorar la validació dels candidats en ser implementades amb el mètode *token slot recognition*.

La implementació conjunta de les estratègies estadístiques que hem analitzat, aplicades a l'extracció automàtica de terminologia, permet identificar els candidats a terme més representatius d'un corpus especialitzat i també endreçar aquests candidats segons la seva major o menor capacitat de ser una unitat terminològica. Aquesta capacitat és determinada per les mesures d'associació lèxica segons la seva dimensió lingüística i estadística.

L'aplicació de les estratègies avaluades en el present treball representa un avenç significatiu respecte al nivell d'identificació de les unitats terminològiques que són presents en un àmbit d'especialitat, també pel que fa a la capacitat de filtratge del nombre de candidats extrets dels corpus i la possibilitat de situar els termes presents en els corpus especialitzats en les posicions inicials d'una llista de resultats, cosa que permet reduir significativament en temps i cost la tasca de validació manual final dels candidats per part d'un especialista.

A més, hem comprovat que l'aplicació combinada d'aquestes estratègies estadístiques ofereix una bona rendibilitat en corpus de diferent mida, de diversos àmbits d'especialitat i de llengües diferents.

Així mateix, és significatiu destacar que la proposta combinada d'extracció de candidats a terme que presentem en aquest treball permet ser utilitzada per a l'extracció de candidats de tipus bigram, trigram i quadrigram amb l'aplicació de la ràtio log-likelihood.

Finalment, hem d'assenyalar que l'ús del mètode *token slot recognition* juntament amb la ràtio log-likelihood, la mesura informació mútua i la prova *t* de Student per a l'extracció de terminologia ofereix prou flexibilitat per a ser integrat en un procés d'extracció més ampli, que inclogui estratègies complementàries de filtratge de candidats.

Capítol 8

TREBALL FUTUR

La línia de treball que dóna continuïtat al present estudi serà l'aplicació de les estratègies estadístiques que hem explorat en el si d'un projecte multilingüe, l'objectiu del qual és obtenir els termes propis d'un àmbit d'especialitat per a ser incorporats en futures obres lexicogràfiques. La dimensió d'aquest projecte ens permetrà avaluar la rendibilitat que obtenen les estratègies d'extracció automàtica de terminologia i de validació final de les unitats resultants, tenint en compte que els resultats obtinguts seran contrastats amb les dades extretes manualment del corpus analitzat.

En el marc d'aquest projecte, el mètode *token slot recognition* i les mesures d'associació lèxica que hem presentat, seran implementats juntament amb estratègies de caràcter lingüístic i híbric. Aquest fet ens permetrà d'ampliar les dades d'anàlisi que hem ofert en el present treball i obtenir un major detall del rendiment de la nostra proposta.

Així mateix, també tenim previst de poder integrar les estratègies estadístiques descrites en el present treball en una eina d'extracció automàtica de terminologia que combini estratègies lingüístiques, estadístiques i híbrides i que serà desenvolupada en codi lliure. Aquesta nova eina permetrà processar corpus especialitzats multilingües procedents de gran diversitat d'àmbits temàtics.

En relació amb l'ús de mesures d'associació lèxica aplicades a la millora dels resultats obtinguts en el procés d'extracció automàtica de terme i a la validació dels candidats a terme per part dels especialistes, tenim previst d'ampliar l'estudi que hem fet amb altres tipus de mesures, algunes de les quals incorporen mètodes híbrids d'extracció.

Finalment, en una altra línia d'estudi, ens plantegem d'incorporar l'anàlisi dels contextos definitoris en què queden circumscrits els termes com a estratègia complementària per a millorar el procés d'extracció automàtica de termes.

Bibliografia

Aguilar, L. (2001). *Lexicografía y terminología aplicadas a la traducción: curso práctico de introducción*, volum 106 de *Materials*. Universitat Autònoma de Barcelona, Servei de Publicacions, Barcelona.

Ahmad, K., Davies, A., Fulford, H., i Rogers, M. (1994). What is a term? the semi-automatic extraction of terms from text. En Snell-Hornby, M., Pöchhacker, F., i Kaindl, K., editors, *Translation Studies: An Interdiscipline*, volum 2, p. 267–278. John Benjamins Publishing Company, Amsterdam.

Alegria, I., Ezeiza, N., Oronoz, M., i Urizar, R. (1999). Extracción automática de terminología a partir de etiquetado y lematización. volum 1, Santiago de Cuba, Cuba.

Amar, M. i David, S. (2001). Evaluation de logiciels d'extraction dans les champs de l'indexation, la traduction et la terminologie. Rapport de recherche établi dans le cadre de l'ARC a3, Agence Universitaire de la Francophonie.

Ananiadou, S. (1988). *Towards a methodology for automatic term recognition*. Tesi doctoral, The University of Manchester, Manchester, Regne Unit.

Ananiadou, S. (1994a). A computational linguistic approach to automatic term recognition. En *Proceedings of the 3rd International Society for Knowledge Organization (ISKO 1994)*, volum 4, p. 134–141, Copenhaguen, Dinamarca. Indeks Verlag.

Ananiadou, S. (1994b). A methodology for automatic term recognition. En *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, volum 2, p. 1034–1038, Kyoto, Japó. Association for Computational Linguistics.

Ananiadou, S., Albert, S., i Schuhmann, D. (2000). Evaluation of automatic term recognition of nuclear receptors from medline. *Genome Informatics*, (11):450–451.

Arppe, A. (1995). Term extraction from unrestricted text. En *Proceedings of the 10th Nordic Conference of Computational Linguistics (NODALIDA 1995)*, Hèlsinki, Finlàndia. Department of General Linguistics.

Aubin, S. i Hamon, T. (2006). Improving term extraction with terminological resources. En *Advances in Natural Language Processing*, volum 4139 de *Lecture Notes in Artificial Intelligence*, p. 380–387. Springer-Verlag Berlin.

Auger, P. (1988). La terminologie au québec et dans le monde, de la naissance à la maturité. En *Actes du 6ème Colloque OLF (Office de la langue française) STQ (Société des traducteurs du Québec) de terminologie : l'ère nouvelle de la terminologie.*, p. 27–59, Mont-real, Canadà. Gouvernement du Québec.

Barité, M. (2013). *Diccionario de organización y representación del conocimiento: clasificación, indización, terminología*. Prodic, Uruguay, 5a. edició.

Baroni, M. i Bernardini, S. (2004). BootCaT: bootstrapping corpora and terms from the web. En *Proceedings of the 4th International Conference on Languages Resources and Evaluation (LREC 2004)*, p. 1313–1316, Elda, Lisboa. European Language Resources Association.

Barrón-Cedeño, A., Sierra, G., Drouin, P., i Ananiadou, S. (2009). An improved automatic term recognition method for spanish. En *Proceedings of the 10th International Conference on Computational Linguis-*

tics and Intelligent Text Processing (CICLING 2009), volum 5449 de *Lecture Notes in Computer Science*, p. 125–136, Mèxic.

Basili, R., De Rossi, G., i Pazienza, M. T. (1997). Inducing terminology for lexical acquisition. En *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP 1997)*, Providence, EUA.

Berthollet, C.-L. (1803). *Essai de statique chimique*. Didot, París.

Bessé, B., Nkwenti-Azeh, B., i Sager, J. C. (1997). Glossary of terms used in terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 4(1):117–156.

Boulaknadel, S., Daille, B., i Aboutajdine, D. (2008). A multi-word term extraction program for arabic language. En *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, p. 1485–1488, Marràqueix, Marroc. European Language Resources Association.

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. En *Proceedings of the 14th Conference on Computational Linguistics (COLING 1992)*, volum 3, p. 977–981, Nantes, França. Association for Computational Linguistics.

Bourigault, D., Gonzalez-Mullier, I., i Gros, C. (1996). LEXTER, a natural language processing tool for terminology extraction. En *Proceedings of the 7th European Association for Lexicography International Congress on Lexicography (EURALEX 1996)*, p. 771–779, Göteborg, Suècia. Göteborg University.

Bourigault, D. i Jacquemin, C. (1999). Term extraction+term clustering: An integrated platform for computer-aided terminology. En *Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics*, p. 15–22, Bergen, Noruega. Association for Computational Linguistics.

Bourigault, D., Jacquemin, C., i L'Homme, M.-C. (2001). *Recent advances in computational terminology*, volum 29 de *Natural language processing*. John Benjamins, Amsterdam.

Bourigault, D. i Slodzian, M. (1999). Pour une terminologie textuelle. En Enguehard, C. i Condamines, A., editors, *Actes des 3èmes Rencontres Terminologie et Intelligence Artificielle (TIA 1999)*, volum 19, p. 29–32, Nantes, França.

Cabré, M. T. (1992). *La terminologia: la teoria, els mètodes, les aplicacions*. Les Naus d'Empúries. Empúries, Barcelona.

Cabré, M. T. (1995). La terminología hoy: concepciones, tendencias y aplicaciones. *Ciência da informação*, 24(3):15.

Cabré, M. T. (1999a). Hacia una teoría comunicativa de la terminología: aspectos metodológicos. En *La Terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos*, volum 11, p. 21–48. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, Barcelona.

Cabré, M. T. (1999b). *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos*. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, Barcelona.

Cabré, M. T. (1999c). ¿Es necesaria una teoría autónoma de la terminología? En *La terminología: representación y comunicación*, p. 93–108. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra, Barcelona.

Cabré, M. T. (2000). Sur la représentation mentale des concepts: bases pour une tentative de modélisation. En Béjoint, H. i Thoiron, P., editors, *Le sens en terminologie*, p. 20–39. Presses Universitaires de Lyon, Lió, França.

Cabré, M. T. (2002). Terminología y lingüística: la teoría de las puertas. *Estudios de lingüística del español*, 16:3.

Cabré, M. T. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology*, 9(2):163–199.

Cabré, M. T. (2010). La teoría comunicativa de la terminología, una aproximación lingüística a los términos. *Revue française de linguistique appliquée*, 14(2):9–15.

Cabré, M. T., Estopà, R., i Vivaldi, J. (2001). Automatic term detection: a review of current systems. En Bourigault, D., Jacquemin, C., i L'Homme, M.-C., editors, *Recent Advances in Computational Terminology*, volum 2 de *Natural language processing*, p. 53–88. John Benjamins, Amsterdam.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., i Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. En *Proceedings of the 3rd International Conference on Language Resources and Evaluation Conference (LREC 2002)*, p. 1934–1940, Las Palmas de Gran Canaria, Espanya. European Language Resources Association.

Chaudiron, S. (2005). Terminologie, ingénierie linguistique et gestion de l'information. *Langages*, 39(157):25–35.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. En *Proceedings of the 1st International Conference on User-Oriented Content-Based Text and Image Handling (RIAO 1988)*, p. 609–623, Cambridge, MA, EUA.

Church, K. W. i Gale, W. A. (1991). Concordances for parallel text. En *Proceedings of the 7th Annual Conference of the Centre for the New Oxford Dictionary and Text Research on Using Corpora*, p. 40–62, Oxford, Anglaterra.

Church, K. W. i Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cohen, W. W., Schapire, R. E., i Singer, Y. (1999). Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270.

Conceição, M. C. (2005). *Concepts, termes et reformulations*. Presses universitaires de Lyon.

Condamines, A. (1994). Terminologie et représentation des connaissances. *Didaskalia: recherches sur la communication et l'apprentissage des sciences et des techniques*, 1(5):35–51.

Condamines, A. (1995). Terminology: New needs, new perspectives. *Terminology*, 2(2):219–238.

Condamines, A. (2005). Linguistique de corpus et terminologie. *Langages*, (1):36–47.

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I., i Flickinger, D. (2002). Multiword expressions: linguistic precision and reusability. En *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, p. 1941–1947, Las Palmas de Gran Canaria, Espanya. European Language Resources Association.

Costa, R. (2006). Texte, terme et contexte. En *Actes des 7es Journées scientifiques du Réseau Lexicologie, Terminologie et Traduction*, p. 79–88, Brussel·les. Editions des Archives Contemporaines et Agence universitaire de la Francophonie.

Cover, T. M. i Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience, 2a. edició.

da Silva, J. F., Dias, G., Guilloré, S., i Pereira Lopes, J. G. (1999). Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. En *Progress in Artificial Intelligence*, volum 1695, p. 113–132. Springer.

Dagan, I. i Church, K. (1994). Termight: identifying and translating technical terminology. En *Proceedings of the 4th Conference on Applied Natural Language Processing*, p. 34–40, Stuttgart, Alemanya. Association for Computational Linguistics.

Daille, B. (1994). *Approche mixte pour l'extraction automatiques de terminologie : statistiques lexicales et filtres linguistiques*. Tesi doctoral, Université Paris 7, París, França.

Daille, B. (1995). *Combined approach for terminology extraction: lexical statistics and linguistic filtering*, volum 5. UCREL Technical Papers, Lancaster, Regne Unit.

Daille, B. (1997). Study and implementation of combined techniques for automatic extraction of terminology. En *The balancing act: combining symbolic and statistical approaches to language*, p. 49–66. Massachusetts Institute of Technology.

Daille, B. (2003). Conceptual structuring through term variations. En *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (MWE 2003)*, volum 18, p. 9–16, Sapporo, Japó. Association for Computational Linguistics.

Daille, B., Gaussier, , i Langé, J.-M. (1998). An evaluation of statistical scores for word association. En Ginzburg, J., Khasidashvili, Z., Vogel, C., Levy, J.-J., i Vallduví, E., editors, *The Tbilisi Symposium on Logic, Language and Computation: Selected Papers*, p. 177–188. Center for the Study of Language and Information.

Daille, B., Habert, B., Jacquemin, C., i Royauté, J. (1996). Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–257.

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29(4):433–447.

David, S. i Plante, P. (1990). Le progiciel TERMINO : de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique de textes. En *Actes du colloque international sur les industries de la langue : perspectives des années 1990*, volum 1, p. 71–88, Montreal, Canadà.

Davis, J. i Goadrich, M. (2006). The relationship between precision-recall and ROC curves. En *Proceedings of the 23rd International Conference on Machine Learning*, p. 233–240, Pittsburgh, Pennsylvania, EUA.

DeGroot, M. H. i Schervish, M. J. (2002). *Probability and Statistics*. Addison-Wesley, Nova York.

Delavigne, V. (2001). *Les mots du nucléaire : contribution socioterminologique à une analyse des discours de vulgarisation*. Tesi doctoral, Université de Rouen, Mont-Saint-Aignan, França.

Depecker, L. (1998). L'ère de la terminologie informationnelle. *Revue française de linguistique appliquée*, 2(3):7–14.

Desmet, I. (1995). *Pour une approche terminologique des sciences sociales et humaines. Les sciences sociales et humaines du travail en portugais et en français*. Tesi doctoral, Université Paris-Nord (Paris 13), París, França.

Dias, G. (2002). *Extraction automatique d'associations lexicales à partir de corpora*. Tesi doctoral, New University of Lisbon, Lisboa, Portugal.

Dias, G. (2003). Multiword unit hybrid extraction. En *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (MWE 2003)*, volum 18, p. 41–48, Sapporo, Japó. Association for Computational Linguistics.

Dias, G., Guilloré, S., Bassano, J.-C., Gabriel, J., i Lopes, J. G. P. (2000). Combining linguistics with statistics for multiword term extraction: A fruitful association? En *Proceedings of Recherche d'Information et ses Applications (RIAO 2000)*, p. 1473–1491, París, França.

Dias, G. i Nunes, S. (2004). Evaluation of different similarity measures for the extraction of multiword units in a reinforcement learning environment. En *Proceedings of the 4th International Conference on Languages Resources and Evaluation (LREC 2004)*, volum 26, p. 1717–1721, Lisboa, Portugal. European Language Resources Association.

Dias, G., Vintar, S., Guilloré, S., i Lopes, J. G. P. (1999). Identifying and integrating terminologically relevant multiword units in the IJS-ELAN slovene-english parallel corpus. En *Proceedings of the 10th Computational Linguisticss in the Netherlands (CLIN 1999)*, Utrecht, Països Baixos.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Diki-Kidiri, M. (2000). Une approche culturelle de la terminologie. *Terminologies nouvelles*, (21):27–31.

Drouin, P. (1997). Une méthodologie d'identification automatique des syntagmes terminologiques : l'apport de la description du non-terme. *Meta*, 42(1):45–54.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Dubuc, R. (1978). *Manuel pratique de terminologie*. Linguatech, Montreal, Canadà.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Dunning, T. E. (1998). *Finding structure in text, genome and other symbolic sequences*. Tesi doctoral, University of Sheffield, Sheffield, Regne Unit.

Earl, L. L. (1970). Experiments in automatic extracting and indexing. *Information Storage and Retrieval*, 6(4):313–330.

Edmundson, H. i Wyllys, R. (1961). Automatic abstracting and indexing-survey and recommendations. *Communications of the Association for Computing Machinery*, 4(5):226–234.

El Hadi, M. (2006). *Terminologie et accès à l'information*. Traité des sciences et techniques de l'information. Hermès science publications, París.

Enguehard, C. i Pantera, L. (1995). Automatic natural acquisition of a terminology. *Journal of quantitative linguistics*, 2(1):27–32.

Estopà, R. (1999). *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. Tesi doctoral, Universitat Pompeu Fabra, Barcelona.

Estopà, R. (2007). Segments no terminològics proposats per un extractor de terminologia com a unitats terminològiques. En *Estudis de lingüística i de lingüística aplicada en honor de M. Teresa Cabré Castellví*, volum 2 de *Documenta Universitaria*, p. 255. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra, Barcelona.

Estopà, R. (2009). Los extractores de terminología: logros y escollos. En Alcina, A., Valero, E., i Rambla, E., editors, *Terminología y sociedad del conocimiento*, p. 117–146. Peter Lang, Alemanya.

Estopà, R., Cabré Castellví, M. T., i Vivaldi, J. (1998). Sistemes d'extracció automàtica de (candidats a) termes: estat de la qüestió. Technical report, Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra, Barcelona.

Evans, D. A. i Zhai, C. (1996). Noun-phrase analysis in unrestricted text for information retrieval. En *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (ACL 1996)*, p. 17–24, Santa Cruz, Califòrnia, EUA. Association for Computational Linguistics.

Everitt, B. S. (1992). *The analysis of contingency tables*. Chapman & Hall, Londres, 2a. edició.

Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*. Tesi doctoral, Universität Stuttgart, Suttgart, Alemanya.

Evert, S. i Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. En *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, p. 188–195, Toulouse, França. Association for Computational Linguistics.

Evert, S. i Krenn, B. (2004a). Association measures.

Evert, S. i Krenn, B. (2004b). Computational approaches to collocations. *Introductory course at the European Summer School on Logic, Language, and Information (ESSLLI 2003)*.

Evert, S. i Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, 19(4):450–466.

Faber Benítez, P. (2009). The cognitive shift in terminology and specialized translation. *MONTI: Monografías de traducción e interpretación*, (1):107–134.

Fabre, C. (1996). *Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations*. Tesi doctoral, Université de Rennes 1, Rennes, França.

Fano, R. M. (1961). *Transmission of Information: A statistical theory of communication*. MIT Press, Nova York.

Faulstich, E. (1998). Principes formels et fonctionnels de la variation en terminologie. *Terminology*, 5(1):93–106.

Felber, H. (1983). Basic principles and methods for the preparation of terminology standards. En *Standardization of technical terminology:*

principles and practices, p. 3–14. American Society for Testing and: Materials, Philadelphia, EUA, 1a. edició.

Felber, H. (1984). *Terminology Manual. General Information Programme and UNISIST*. International Information Centre for Terminology, Paris: UNESCO: Infoterm.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. En *Studies in linguistic analysis.Special volume of the Philological Society*, p. 1–32. The Philological Society, Oxford.

Foo, J. (2011). Exploring termhood using language models. En *Proceedings of the Workshop on Creation, Harmonization and Application of Terminology Resources (CHAT 2011)*, NEALT Proceedings Series; Vol. 12, p. 32–35, Riga, Letònia. Northern European Association for Language Technology.

Foo, J. (2012). *Computational Terminology: Exploring Bilingual and Monolingual Term Extraction*. Tesi doctoral, Linköping University, Linköping, Suècia.

Frantzi, K. i Ananiadou, S. (1996a). A hybrid approach to term recognition. En *Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP-IA 1996)*, volum 1, p. 93–98, Moncton, Canadà.

Frantzi, K. i Ananiadou, S. (1997). Automatic term recognition using contextual cues. En *Proceedings of 3rd Delos Workshop*, Zurich, Suïssa.

Frantzi, K. i Ananiadou, S. (1999). The c-value/NC-value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–179.

Frantzi, K., Ananiadou, S., i Mima, H. (2000). Automatic recognition of multi-word terms: the c-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Frantzi, K. T. i Ananiadou, S. (1996b). Extracting nested collocations. En *Proceedings of the 16th Conference on Computational Linguistics (COLING 1996)*, volum 1, p. 41–46, Copenhaguen, Dinamarca.

Gaizauskas, R., Demetriou, G., i Humphreys, K. (2000). Term recognition and classification in biological science journal articles. En *Proceedings of the Computational Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on Natural Language Processing (NLP 2000)*, p. 37–44, Patras, Grècia.

Gambier, Y. (1987). Problèmes terminologiques des pluies acides: pour une socio-terminologie. *Meta : Journal des traducteurs*, 32(3):314–320.

Gambier, Y. (1991). Présupposés de la terminologie: vers une remise en cause. *Cahiers de linguistique sociale*, 18:31–58.

Gambier, Y. (1993). Implications épistémologiques et méthodologiques de la socioterminologie. En *Actes du 15ème Congrès International des Linguistes*, p. 14–15, Quebec, Canadà. Presses de l'Université Laval.

Gaudin, F. (1993). *Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles*. Publications de l'Université de Rouen.

Gaudin, F. (2003). *Socioterminologie: une approche sociolinguistique de la terminologie*. Champs Linguistiques. De Boeck Supérieur, Brussel·les.

Gaudin, F. i Alexandru, C. (2005). Les contextes : à la source du terme ? En *Actes du colloque Mots, termes et contextes*, volum 7, p. 59–69, Brussel·les, Bèlgica.

Gojun, A., Heid, U., Weissbach, B., Loth, C., i Mingers, I. (2012). Adapting and evaluating a generic term extraction tool. En *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, p. 651–656, Istanbul, Turquia. European Language Resources Association.

Gómez González-Jover, A. (2007). *Terminografía, lenguajes profesionales y mediación interlingüística: aplicación metodológica al léxico especializado de la industria del calzado y las industrias afines*. Tesi doctoral, Universitat d'Alacant, Alacant.

Ha, L. A. (2007). *Advances in automatic terminology processing: methodology and application in focus*. Tesi doctoral, University of Wolverhampton, Wolverhampton, Regne Unit.

Heid, U. (1999). Extracting terminologically relevant collocations from german technical texts. En *Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE 1999)*, p. 241–255, Innsbruck, Àustria.

Heid, U. (2006). Extracting term candidates from recursively chunked text. *Terminology, Computing and Translation. Tübingen: Gunter Narr*, p. 97–116.

Heid, U. i McNaught, J. (1991). EUROTRA-7 study: Feasibility and project definition study on the reusability of lexical and terminological resources in computerised applications. Final report. CEC-DG XIII.

Hoffmann, L. (1998). *Llenguatges d'especialitat: selecció de textos*. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra, Barcelona.

Hovy, E., Gerber, L., Hermjakob, U., Junk, M., i Lin, C.-Y. (2000). Question answering in webclopedia. En *Proceedings of the 9th Text Retrieval Conference (TREC 2000)*, p. 655–532, Gaithersburg, Maryland, EUA. Department of Commerce. National Institute of Standards and Technology.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, p. 216–223, Sapporo, Japó. Association for Computational Linguistics.

Huo, W. (2012). *Automatic multi-word term extraction and its application to web-page summarization*. Tesi doctoral, University of Guelph, Guelph, Ontario, Canadà.

ISO (2000). ISO 1087-1:2000 terminology work. vocabulary. part 1: Theory and application.

ISO (2009). ISO 704:2009 terminology work. principles and methods.

ISO (2012). TC 37 terminology and other language and content resources.

Ittoo, A. i Bouma, G. (2013). Term extraction from sparse, ungrammatical domain-specific documents. *Expert Systems with Applications*, 40(7):2530–2540.

Jaccard, P. (1901). *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*, volum 37. Bulletin de la Société vaudoise des sciences naturelles, Lausanne.

Jacquemin, C. (1994). FASTR: a unification-based front-end to automatic indexing. En *Proceedings of the 4th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) (RIAO 1994)*, volum 2, p. 34–47, Nova York, EUA. Rockfeller University Press.

Jacquemin, C. (1997). Variation terminologique : reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Technical report, Université de Nantes, Nantes, França.

Jacquemin, C. (1999). Syntagmatic and paradigmatic representations of term variation. En *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, p. 341–348, College Park, Maryland, EUA. Association for Computational Linguistics.

Jacquemin, C. (2001). *Spotting and discovering terms through natural language processing*. The MIT Press.

Jacquemin, C., Klavans, J. L., i Tzoukermann, E. (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. En *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL 1997)*, p. 24–31, Madrid, Spain. Association for Computational Linguistics.

Justeson, J. S. i Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Kageura, K. (2002). *The dynamics of terminology: a descriptive theory of term formation and terminological growth*, volum 5 de *Terminology and Lexicography Research and Practice*. John Benjamins Publishing Company.

Kageura, K. i Umino, B. (1996). Methods of automatic term recognition: a review. *Terminology*, 3(2):259–289.

Kennedy, G. (1998). *An introduction to corpus linguistics*. Studies in Language and Linguistics. Addison-Wesley-Longman, Londres, Nova York.

Kerremans, K., Temmerman, R., i Zhao, G. (2005). Terminology and knowledge engineering in fraud detection. En *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering (TKE 2005)*, Copenhaguen, Dinamarca.

Kocourek, R. (2001). Le terme et sa définition. En *Essais de linguistique française et anglaise. Mots et termes, sens et textes.*, volum 48 de *Bibliothèque de l'Information Grammaticale*, p. 271–297. Peeters Publishers, Lovaina, París.

Kohli, S. (2006). *Introducing an object oriented design to the ngram statistics package*. Tesi doctoral, University of Minnesota, Duluth, Minnesota, EUA.

Korkontzelos, I., Klapaftis, I. P., i Manandhar, S. (2008). Reviewing and evaluating automatic term recognition techniques. En *Advances in Natural Language Processing*, volum 5221 de *Lecture Notes in Computer Science*, p. 248–259. Springer.

Krauthammer, M. i Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of biomedical informatics*, 37(6):512–526.

Krenn, B. i Evert, S. (2001). Can we do better than frequency? a case study on extracting PP-verb collocations. En *Proceedings of the ACL Workshop on Collocations*, p. 39–46, Toulouse, França.

Lauriston, A. (1995). Criteria for measuring term recognition. En *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1995)*, p. 17–22, Dublin, Irlanda. Morgan Kaufmann Publishers Inc.

Lauriston, A. (1996). *Automatic term recognition: performance of linguistic and statistical techniques*. Tesi doctoral, University of Manchester, Manchester, Regne Unit.

Lavoisier, A. (1793). *Traité élémentaire de Chimie*. Chez Cuchet, Libraire, París.

Linné, C. v. (1736). *Fundamenta botanica*. Salomon Schouten, Amsterdam, 1a. edició.

Loginova, E., Gojun, A., Blancafort, H., Guégan, M., Gornostay, T., i Heid, U. (2012). Reference lists for the evaluation of term extraction tools. En *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)*, p. 177–192, Madrid, Spain.

Lyse, G. I. i Andersen, G. (2012). Collocations and statistical analysis of n-grams. En *Exploring newspaper language: using the web to create and investigate a large corpus of modern Norwegian*, volum 49 de *Studies in Corpus Linguistics*, p. 79–110.

Manning, C. D., Raghavan, P., i Schütze, H. (2008). *Introduction to information retrieval*, volum 1. Cambridge University Press, Cambridge.

Manning, C. D. i Schütze, H. (2003). *Foundations of statistical natural language processing*. MIT Press, 6a. edició.

Maynard, D. (2000). *Term recognition using combined knowledge sources*. Tesi doctoral, Manchester Metropolitan University, Manchester, Regne Unit.

Maynard, D. i Ananiadou, S. (1999). Identifying contextual information for multi-word term extraction. En *Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE 1999)*, p. 212–221, Innsbruck, Àustria. TermNet.

Maynard, D. i Ananiadou, S. (2000a). Identifying terms by their family and friends. En *Proceedings of the 18th Conference on Computational Linguistics (COLING 2000)*, volum 1, p. 530–536, Saarbrücken, Alemanya. Association for Computational Linguistics.

Maynard, D. i Ananiadou, S. (2000b). Trucks: a model for automatic multi-word term recognition. *Journal of Natural Language Processing*.

McEnery, T., Langé, J.-M., Oakes, M., i Véronis, J. (1997). The exploitation of multilingual annotated corpora for term extraction. En *Corpus annotation: linguistic information from computer text corpora*, p. 220–230. Addison Wesley Longman, Boston, MA, EUA.

McInnes, B. T. (2004). *Extending the log likelihood measure to improve collocation identification*. Tesi doctoral, University of Minnesota, Minneapolis, MN, EUA.

Merkel, M. i Andersson, M. (2000). Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. En *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) (RIAO 2000)*, p. 737–746, París, França.

Merkel, M. i Foo, J. (2007). Terminology extraction and term ranking for standardizing term banks. En *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, p. 349–354, Tartu, Estònia. University of Tartu.

Meyer, I., Bowker, L., i Eck, K. (1992). Cogniterm: an experiment in building ? terminological knowledge base. En *Proceedings of the 5th European Association for Lexicography International Congress*, p. 159–172, Tampere, Finlàndia. Department of Translation Studies, University of Tampere.

Meyer, I. i Mackintosh, K. (2000). When terms move into our everyday lives: An overview of de-terminologization. *Terminology*, 6(1):111–138.

Milios, E., Zhang, Y., He, B., i Dong, L. (2003). Automatic term extraction and document similarity in special text corpora. En *Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics*, p. 275–284, Halifax, NS, Canadà.

Miller, G. A. (1995). WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mima, H., Ananiadou, S., i Nenadi?, G. (2001). The ATRACT workbench: automatic term recognition and clustering for terms. En *Text, Speech and Dialogue*, volum 2166 de *Lecture Notes in Computer Science*, p. 126–133. Springer Berlin Heidelberg.

Mohanan, K. P. (1986). *The theory of lexical phonology*. Studies in Natural Language and Linguistic Theory. Springer.

Montanes, E., Diaz, I., Ranilla, J., Combarro, E. F., i Fernandez, J. (2005). Scoring and selecting terms for text categorization. *Intelligent Systems, IEEE*, 20(3):40–47.

Mustafa el Hadi, W. (2005). Indexation humaine et indexation automatisée : la place du terme et de son environnement. En *Actes du colloque*

Mots, termes et contextes dans les 7es Journées Scientifiques du Réseau LTT, p. 157–167, Brussel·les, Bèlgica.

Nakagawa, H. i Mori, T. (2002). A simple but powerful automatic term extraction method. En *Proceedings of the 2nd International Workshop on Computational Terminology (COMPUTERM 2002)*, volum 14, p. 29–35, Taipei, Taiwan. Association for Computational Linguistics.

Nakagawa, H. i Mori, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

Naulleau, (1998). *Apprentissage et filtrage syntactico-sémantique de syntagmes nominaux pertinents pour la recherche documentaire*. Tesi doctoral, Université Paris 13, Villetaneuse, França.

Nenadic, G., Ananiadou, S., i McNaught, J. (2004). Enhancing automatic term recognition through recognition of variation. En *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, p. 604–610, Ginebra, Suïssa.

Nenadic, G., Spasic, I., i Ananiadou, S. (2005). Mining biomedical abstracts: what's in a term? En *Natural Language Processing (IJCNLP 2004)*, volum 3248 de *Lecture Notes in Computer Science*, p. 797–806.

Nkwenti-Azeh, B. (1994). Positional and combinational characteristics of terms: consequences for corpus-based terminography. *Terminology*, 1(1):61–95.

Oliver, A., Moré, J., i Climent, S. (2007). *Traducció i tecnologies*, volum 116 de *Manuals*. Editorial UOC, Barcelona.

Pal, S., Kumar Naskar, S., Pecina, P., Bandyopadhyay, S., i Way, A. (2010). Handling named entities and compound verbs in phrase-based statistical machine translation. En *Proceedings of the Workshop on Multiword Expressions: from teory to applications (MWE 2010)*, Pequín, Xina.

Pazienza, M. T., Pennacchiotti, M., i Zanzotto, F. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. En *Knowledge Mining*, volum 185 de *Studies in Fuzziness and Soft Computing*, p. 255–279. Springer Berlin Heidelberg.

Pearson, J. (1998). *Terms in context*, volum 1 de *Studies in Corpus Linguistics*. John Benjamins Publishing Company.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1):137–158.

Pecina, P. i Schlesinger, P. (2006). Combining association measures for collocation extraction. En *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, p. 651–658, Sidney, Austràlia. Association for Computational Linguistics.

Pereira, R., Crocker, P., i Dias, G. (2004). A parallel multikey quicksort algorithm for mining multiword units. En *Proceedings of the Workshop on Methodologies and Evaluation of Multiword Units in Real-world Application*, Lisboa, Portugal.

Peñas, A., Gonzalo, J., i Verdejo, F. (2001). Cross-language information access through phrase browsing. En *Proceedings of the 6th International Workshop on Applications of Natural Language to Information Systems*, volum 3, p. 121–130, Madrid, Spain.

Piao, S. S. i McEnery, T. (2001). Multi-word unit alignment in english-chinese parallel corpora. En *Proceedings of the Corpus Linguistics Conference*, volum 13, p. 466–475, Lancaster, Regne Unit.

Piao, S. S., Rayson, P., Archer, D., i McEnery, T. (2005). Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech & Language*, 19(4):378–397.

Quasthoff, U. i Wolff, C. (2002). The poisson collocation measure and its applications. En *Proceedings of the 2nd International Workshop on Computational Approaches to Collocations*, Viena, Àustria.

Roche, M., Azé, J., Kodratoff, Y., i Sebag, M. (2004). Learning interestingness measures in terminology extraction. a ROC-based approach. En *Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence*, p. 81–88, València, Espanya.

Rocheteau, J. i Daille, B. (2011). TTC TermSuite: a UIMA application for multilingual terminology extraction from comparable corpora. En *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, p. 9–12, Chiang Mai, Tailàndia.

Rondeau, G. (1984). *Introduction à la terminologie*. Gaétan Morin, Quebec, 2a. edició.

Sager, J. C. (1990). *A practical course in terminology processing*. John Benjamins Publishing Company.

Sager, J. C. (1998). In search of a foundation: towards a theory of the term. *Terminology*, 5(1):41–57.

Sager, J. C. (2000). Pour une approche fonctionnelle de la terminologie. En *Le sens en terminologie*, p. 40–60. Presses Universitaires Lyon, Lió.

Sager, Juan C., Dungworth, David, i McDonald, Peter F. (1980). *English special languages: principles and practice in science and technology*. Brandstetter, Wiesbaden.

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley.

Schmidt, F. (2001). *A comparison of different approaches to multi-word term acquisition*. Tesi doctoral, Universität München, München, Alemanya.

Selkirk, E. O. (1982). *The syntax of words*. Linguistic Inquiry Monographs. MIT Press, Cambridge.

Slodzian, M. (1993). La VGTT (vienna general theory of terminology) et la conception scientifique du monde. *Le langage et l'homme*, 28(4):223–232.

Slodzian, M. (1995). Comment revisiter la doctrine terminologique aujourd'hui ? *La banque des mots*, p. 11–18.

Slodzian, M. (2000). L'émergence d'une terminologie textuelle et le retour du sens. *Le sens en terminologie*, p. 61–85.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Smadja, F., McKeown, K. R., i Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1–38.

Tartier, A. (2001). Méthodes d'analyse automatique de l'évolution terminologique au travers des variations repérées dans les corpus diachroniques. En *Actes des 4èmes Rencontres Terminologie et Intelligence Artificielle (TIA 2001)*, p. 191–200, Nancy, França. Vandoeuvre lès Nancy.

Temmerman, R. (2000). *Towards new ways of terminology description: the sociocognitive approach*, volum 3 de *Terminology and Lexicography Research and Practice*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Temmerman, R. i Kerremans, K. (2003). Termontography: ontology building and the sociocognitive approach to terminology description. En *Proceedings of the 17th International Congress of Linguists*, volum 1, Praga, República Txeca.

Thanopoulos, A., Fakotakis, N., i Kokkinakis, G. (2002). Comparative evaluation of collocation extraction metrics. En *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volum 2, p. 620–625, Las Palmas de Gran Canaria, Espanya.

Tsay, J.-J. i Wang, J.-D. (1999). Term selection with distributional clustering for chinese text categorization using n-grams. En *Proceedings of the 12th Conference on Computational Linguistics (ROCLING 1999)*, volum 2, p. 151–170, Hsinchu, Taiwan.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. En *Proceedings of the 12th European Conference on Machine Learning (ECML 2001)*, volum 2167 de *Lecture Notes in Computer Science*, p. 491–502, Freiburg, Alemanya. Springer-Verlag.

UNE (2009). UNE-ISO 1087-1:2009 treball terminològic. vocabulari. part i: Teoria i aplicació.

Utiyama, M., Murata, M., i Isahara, H. (2000). Using author keywords for automatic term recognition. *Japanese Term Extraction. Special issue of Terminology*, 6(2):313–326.

Velardi, P., Missikoff, M., i Basili, R. (2001). Identification of relevant terms to support the construction of domain ontologies. En *Proceedings of the Workshop on Human Language Technology and Knowledge Management*, p. 1–8, Toulouse, França. Association for Computational Linguistics.

Vivaldi, J. (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Tesi doctoral, Universitat Politècnica de Catalunya, Barcelona.

Vivaldi, J., Màrquez, L., i Rodríguez, H. (2001). Improving term extraction by system combination using boosting. *Machine Learning: ECML 2001*, p. 515–526.

Vivaldi, J. i Rodríguez, H. (2007). Evaluation of terms and term extraction systems: a practical approach. *Terminology*, 13(2):225–248.

Vossen, P. (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers.

Vu, T., Aw, A. T., i Zhang, M. (2008). Term extraction through unithood and termhood unification. En *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, volum 1, p. 631–636, Hyderabad, Índia.

Vàzquez, M. i Oliver, A. (2007). Anàlisi de tècniques estadístiques d'extracció automàtica de termes. Treball de recerca, Universitat Pompeu Fabra, Barcelona.

Wermter, J. (2009). *Collocation and term extraction using linguistically enhanced statistical methods*. Tesi doctoral, Universität Jena, Jena, Alemanya.

Wermter, J. i Hahn, U. (2004). Collocation extraction based on modifiability statistics. En *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, p. 980–986, Ginebra, Suïssa. Association for Computational Linguistics.

Wermter, J. i Hahn, U. (2005a). Massive biomedical term discovery. En *Discovery Science*, volum 3735 de *Lecture Notes in Computer Science*, p. 281–293. Springer.

Wermter, J. i Hahn, U. (2005b). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. En *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, p. 843–850, Vancouver, Canadà. Association for Computational Linguistics.

Wermter, J. i Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures

for collocation and term extraction. En *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, p. 785–792, Sydney, Austràlia. Association for Computational Linguistics.

Witschel, H.-F. (2005). Terminology extraction and automatic indexing-comparison and qualitative evaluation of methods. En *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering (TKE 2005)*, Copenhaguen, Dinamarca.

Wüster, E. (1931). *Internationale Sprachnormung in der Technik: besonders in der Elektrotechnik. Die nationale Sprachnormung und ihre Verallgemeinerung*. Tesi doctoral, Universität Stuttgart, Stuttgart, Alemanya.

Wüster, E. (1968). *The machine tool: an interlingual dictionary of basic concepts: comprising an alphabetical dictionary and a classified vocabulary with definitions and illustrations*. Technical Press, Londres, 1a edició.

Wüster, E. (1979a). *Einführung in die allgemeine Terminologielehre und erminologische Lexikographie*. The Copenhagen School of Economics, Viena/Nova York, springer edició.

Wüster, E. (1979b). *Introduction to the general theory of terminology and terminological lexicography*. Springer, Viena/Nova York.

Yoshikane, F., Tsuji, K., Kageura, K., i Jacquemin, C. (1999). Detecting japanese term variation in textual corpus. En *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages (IRAL 1999)*, p. 97–108, Taipei, Taiwan. Academia Sinica.

Zanzotto, F. M. (2002). *L'estrazione della terminologia come strumento per la modellazione di domini conoscitivi*. Tesi doctoral, Università degli Studi di Roma "Tor Vergata", Roma, Itàlia.

Zervanou, K. i McNaught, J. (2004). A domain-independent approach to IE rule development. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, p. 745–748, Lisboa, Portugal.

Zhang, C., Niu, Z., Jiang, P., i Fu, H. (2012). Domain-specific term extraction from free texts. En *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012)*, p. 1290–1293, Sichuan, Xina.

ANNEXOS

Apèndix A

ANNEX I: MESURES D'ASSOCIACIÓ LÈXICA

En el present annex recollim els resultats obtinguts per les mesures d'associació lèxica amb relació al nombre de candidats a terme que s'ha de revisar manualment en cada corpus especialitzat. Aquests resultats estan organitzats en tres franges percentuals: 25%, 50% i 75% dels resultats. En cada franja destaquem el percentatge més alt i més baix de candidats a terme que s'ha de revisar manualment per a arribar a identificar els termes que són presents en cada corpus especialitzat.

Taula A.1: Distribució dels termes en els corpus especialitzats (I).

Coeficient Dice							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	**12,1%**	**19**	**30,5%**	**48**	63%	99
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	16,9%	62	32,3%	118	72%	263
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	15,6%	15	32,2%	31	**61,4%**	**59**
Economia esp.	152 termes	25%	38	50%	76	75%	114
IULA	414 CT	18,5%	77	48,3%	200	74,3%	308
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	25%	20	58,7%	47	**81,2%**	**65**
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	**25,5%**	**22**	**59,3%**	**51**	62,7%	54
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	14,4%	115	40,3%	320	68,6%	545
Ràtio log-likelihood							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	**17,8%**	**28**	37,5%	59	63,6%	100
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	**9,5%**	**35**	38%	139	**74,5%**	**272**
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	15,6%	15	40,6%	39	69,7%	67
Economia esp.	152 termes	25%	38	50%	76	75%	114
IULA	414 CT	11,1%	46	38,8%	161	70%	290
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	13,7%	11	**61,2%**	**49**	71,2%	57
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	16,2%	14	53,4%	46	**55,8%**	**48**
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	11,3%	90	**32,8%**	**261**	64,1%	509

Taula A.2: Distribució dels termes en els corpus especialitzats (II).

Ràtio odds							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	17,8%	28	**38,8%**	**61**	66,8%	105
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	21,6%	79	52%	190	73,9%	270
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	**14,5%**	**14**	41,6%	40	69,7%	67
Economia esp.	152 termes	25%	38	50%	76	75%	114
IULA	414 CT	20%	83	48,3%	200	74,8%	310
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	23,7%	19	**61,2%**	**49**	**77,5%**	**62**
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	**24,4%**	**21**	56,9%	49	**61,6%**	**53**
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	17,1%	136	42,4%	337	72,7%	578
Coeficient PHI^2							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	17,8%	28	**36,3%**	**57**	66,8%	103
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	19,4%	71	41,9%	153	73,4%	268
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	17,7%	17	40,6%	39	68,7%	66
Economia esp.	152 termes	25%	38	50%	76	75%	114
IULA	414 CT	20,5%	85	42,7%	177	71,7%	297
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	21,2%	17	**58,7%**	**47**	**75%**	**60**
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	**27,9%**	**24**	56,9%	49	**59,3%**	**51**
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	**15,4%**	**123**	43,1%	343	69%	548

Taula A.3: Distribució dels termes en els corpus especialitzats (III).

Mesura pointwise mutual information							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	23,5%	37	47,7%	75	72,6%	114
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	27,9%	102	53,6%	196	**76,7%**	**280**
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	**16,6%**	**16**	**42,7%**	**41**	**69,7%**	**67**
Economia esp.	152 termes	25%	38	50%	76	75%	114
IULA	414 CT	28%	116	57,9%	240	76,3%	316
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	**35%**	**28**	**63,7%**	**51**	73,7%	59
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	30,2%	26	58,1%	50	69,9%	55
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	21,4%	170	44,7%	268	73,9%	587
Mesura Poisson-Stirling							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	10,8%	17	**34,3%**	**54**	65,6%	103
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	**10,1%**	**37**	35,3%	129	73,4%	268
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	14,5%	14	37,5%	36	69,7%	67
Economia esp.	152 termes	25%	38	50%	76	75%	114
IULA	414 CT	11,8%	49	38,6%	160	71,9%	298
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	12,5%	10	**58,7%**	**47**	**73,7%**	**59**
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	**16,2%**	**14**	56,9%	49	**63,9%**	**55**
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	12,7%	101	34,6%	**275**	68,1%	541

Taula A.4: Distribució dels termes en els corpus especialitzats (IV).

Mesura informació mútua							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	10,1%	16	34,3%	54	64,9%	102
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	**9,5%**	**35**	38%	139	**74,7%**	**273**
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	15,6%	15	40,6%	39	69,7%	67
Economia esp.	152 termes	25%	46	50%	76	75%	114
IULA	414 CT	11,8%	49	38,8%	161	70%	290
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	13,7%	11	**61,2%**	**49**	70%	56
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	**17,4%**	**15**	53,4%	46	**55,8%**	**48**
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	11,5%	92	**32,8%**	**261**	64,2%	510
Mesura *t* Student							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	10,8%	17	29,3%	46	63,6%	100
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	10,4%	38	**28,4%**	**104**	73,6%	269
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	12,5%	12	37,5%	36	66,6%	64
Economia esp.	152 termes	25%	38	50%	76	75%	114
IULA	414 CT	**10,3%**	**43**	31,6%	131	72,2%	299
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	13,7%	11	**58,7%**	**47**	**75%**	**60**
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	**17,4%**	**15**	50%	43	**54,6%**	**47**
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	12,8%	102	28,7%	228	65,4%	520

Taula A.5: Distribució dels termes en els corpus especialitzats (V).

Prova khi quadrat de Pearson							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	17,8%	28	**36,3%**	**57**	65,6%	103
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	19,4%	71	41,9%	153	73,4%	268
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	17,7%	17	40,6%	39	68,7%	66
Economia esp.	152 termes	25%	38	50%	76	75%	114
IULA	414 CT	20,5%	85	42,7%	177	71,7%	297
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	21,2%	17	**58,7%**	**47**	**75%**	**60**
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	**27,9%**	**24**	56,9%	49	**59,3%**	**51**
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	**15,4%**	**123**	43,3%	344	69%	548
Coeficient Jaccard							
Economia esp.	56 termes	25%	14	50%	28	75%	42
JRC	157 CT	17,8%	28	37,5%	59	63,6%	100
Economia ang.	83 termes	25%	21	50%	42	75%	62
JRC	365 CT	16,9%	62	32,3%	118	72%	263
Economia fr.	39 termes	25%	10	50%	20	75%	29
JRC	96 CT	15,6%	15	**32,2%**	**31**	**61,4%**	**59**
Economia esp.	152 termes	25%	38	50%	76	75%	114
IULA	414 CT	18,5%	77	48,3%	200	74,3%	308
Serv. soc. cat.	27 termes	25%	7	50%	14	75%	20
Termcat	80 CT	25%	20	58,7%	47	**81,2%**	**65**
Serv. soc. esp.	35 termes	25%	9	50%	18	75%	20
Termcat	86 CT	**25,5%**	**22**	**59,3%**	**51**	62,7%	54
Medicina esp.	159 termes	25%	40	50%	80	75%	119
IULA	794 CT	**14,4%**	**115**	38,1%	303	68,6%	545