
Aprenentatge automàtic de classes sintàctiques verbals

Laia Mayol Toll

Treball de recerca
Director: Dr. Toni Badia

Doctorat en Ciència Cognitiva i Llenguatge
Bienni 2002-2004
Universitat Pompeu Fabra
Barcelona, juny del 2004

Agraïments

Vull agrair al Toni Badia que hagi acceptat dirigir aquest treball i que m'hagi donat sempre bons consells per tirar-lo endavant. També vull donar gràcies a l'Enric Vallduví pel suport continu que m'ha donat des que vaig acabar la carrera.

A la Gemma Boleda, moltíssimes gràcies per guiar-me durant tots els mesos en què he elaborat aquest treball. Gràcies per tots els consells, per la lectura crítica de bona part del treball i per les nombroses estones que ha dedicat a ajudar-me amb paciència i entusiasme.

A tots els companys del Glicom -Gemma Boleda, Stefan Bott, Beto Boullosa, Judith Domingo, Àngel Gil, Maite Melero, Toni Oliver, Martí Quixal, Ana Ruggia i Oriol Valentín- gràcies per totes les converses, els ànims i els cafès compartits a mig matí.

Als meus pares i al meu germà, gràcies per animar-me i ajudar-me sempre. A la Mercè i al Joan, gràcies per riure amb mi (i de mi) quan cal.

Dóno les gràcies a l'Institut d'Estudis Catalans i al Departament de Traducció i Filologia de la Universitat Pompeu Fabra per haver-me deixat fer servir un fragment del corpus CTILC, sense el qual no hauria pogut fer aquest treball de recerca.

Aquest treball ha estat realitzat amb el suport del Departament d'Universitats, Recerca i Societat de la Informació (beca 2003FI-00867).

Llistat de continguts

Agraïments	ii
Llistat de continguts	iii
Llistat de figures	v
Llistat de taules	vi
Abreviatures	vii
1 Introducció	1
2 Revisió bibliogràfica	3
2.1 Introducció	3
2.2 El verb	3
2.2.1 El sintagma verbal dins l'oració	3
2.2.2 Alternances	7
2.2.3 La partícula <i>se</i>	11
2.2.4 Algunes propostes de classificació	12
2.2.5 Conclusions	14
2.3 L'adquisició lèxica	15
2.3.1 Treballs previs d'adquisició de la subcategorització dels verbs	15
2.3.2 Els primers treballs: compilació de marcs de subcategorització	16
2.3.3 Classificació automàtica	18
2.3.4 Classificació semàntica	21
2.3.5 Conclusions	22
3 Experiments amb mètode supervisat	23
3.1 Introducció	23
3.2 Materials i mètodes	23
3.2.1 El corpus	23
3.2.2 Classificació de verbs	25
3.3 Definició de trets	26
3.4 Experiments i resultats	31
3.5 Anàlisi de resultats	35
3.5.1 Anàlisi d'errors	35
3.5.2 Avaluació	38
3.6 Conclusions	39
4 Experiments amb <i>clustering</i>	41
4.1 Introducció	41
4.2 Materials i mètodes	41
4.2.1 Classificació dels verbs	41
4.2.2 Definició de trets	44
4.2.3 Paràmetres experimentals	44
4.3 Resultats	46
4.3.1 Solució en tres <i>clusters</i>	46
4.3.2 Anàlisi d'errors	50
4.3.3 Solució en quatre <i>clusters</i>	53
4.4 Conclusions	55

5	Experiments addicionals	57
5.1	Introducció	57
5.2	<i>Clustering</i> amb un corpus etiquetat automàticament	57
5.2.1	La <i>Constraint Grammar</i>	57
5.2.2	Solució en tres <i>clusters</i>	59
5.2.3	Solució en quatre <i>clusters</i>	62
5.2.4	Conclusions	63
5.3	Cap a una adquisició dels complement preposicionals	64
5.3.1	Introducció	64
5.3.2	Trets	66
5.3.3	Resultats	68
5.3.4	Conclusions	69
6	Conclusions i treball futur	71
6.1	Conclusions	71
6.2	Treball futur	73
7	Annex A. Regles extretes fent servir la metodologia <i>cross-validation</i>	74
8	Annex B. Solució de <i>clustering</i>	77
9	Annex C. Llexicó	79
	Bibliografia	98

Llistat de figures

2.1	Exemples de verbs inergatiu, inacusatiu i <i>object-drop</i> (Merlo i Stevenson, 2001, pàgina 374)	19
3.1	Exemple de l' <i>input</i> per a Ripper	23
3.2	Extracte del CTILC	25
3.3	Exemple de la codificació verbal	25
3.4	Valors dels trets per als verbs <i>contemplar</i> i <i>xisclar</i>	31
3.5	Resultats de Ripper amb entrenament i avaluació	32
3.6	Resultats de Ripper amb <i>cross-validation</i>	34
3.7	Verbs mal classificats en el GA	35
4.1	Classificació manual de 200 verbs	43
4.2	Solució en 3 <i>clusters</i> . Els colors representen, de més clar a més fosc, transitius, VASE i intransitius	47
4.3	Característiques dels centroides dels 3 <i>clusters</i>	50
4.4	Solució en 4 <i>clusters</i> . Els colors representen, de més clar a més fosc, transitius, VASE i intransitius	54
5.1	Extracte del CIEC etiquetat amb la CatCG	58
5.2	Solució en 3 i 4 <i>clusters</i> amb el CCG. Els colors representen, de més clar a més fosc, transitius, VASE i intransitius	60
5.3	Classificació manual dels 212 verbs	66
5.4	Solució en 2 <i>clusters</i> . Els colors representen, de més clar a més fosc, transitius, règim verbal i intransitius purs	68

Llistat de taules

3.1	Distribució de textos literaris al CTILC, segons el gènere	24
3.2	Distribució de textos no literaris al CTILC, segons l'àrea	24
3.3	Percentatge d'ocurrències seguides de DN i d'ocurrències VNP + DN. Els cinc primers verbs són transitius i els cinc darrers són intransitius	30
3.4	Valors dels trets per als transitius i per als intransitius	32
3.5	Taxes d'error eliminant trets	34
4.1	Valors dels trets per als transitius, VASE i intransitius	44
4.2	Índexs de cobertura, precisió i coeficient F	48
4.3	Preposicions més freqüents que segueixen els verbs VASE	49
4.4	Verbs mal classificats	51
4.5	Ocurrències intransitives de verbs de la classe VASE	52
4.6	Exemple dels verbs dels <i>clusters</i> 1 i 3 de la solució en 4 <i>clusters</i>	53
5.1	Valors dels trets amb el corpus etiquetat per la CatCG	59
5.2	Índexs de cobertura, precisió i coeficient F amb el CCG	60
5.3	Índexs de cobertura, precisió i coeficient F amb quatre <i>clusters</i>	62
5.4	Exemple de verbs dels <i>clusters</i> 0 i 2 de la solució en 4 <i>clusters</i> amb el CCG	63
5.5	Verbs mal classificats amb les diverses anotacions	63
5.6	Valors dels trets per als verbs de règim verbal, intransitius i transitius	68
5.7	Índexs de cobertura, precisió i coeficient F	68
5.8	Verbs intransitius i de règim mal classificats	70
5.9	Preposicions més freqüents que segueixen els verbs del <i>cluster</i> 0	70

Abreviatures

CatCG	<i>Constraint Grammar</i> del català
CCG	Corpus CTILC etiquetat amb l' anotació de la CatCG
CIEC	Corpus CTILC etiquetat amb l' anotació de l' IEC
CTILC	Corpus Textual Informatitzat de la Llengua Catalana
GA	Grup d'avaluació
GE	Grup d'entrenament
LC	Lingüística Computacional
MSC	Marc de Subcategorització
PLN	Processament del Llenguatge Natural
SN	Sintagma nominal
SP	Sintagma preposicional
SV	Sintagma verbal
VASE	Verbs d'alternança amb <i>se</i>
VNP	Verb en forma no personal

Introducció

Aquest treball pretén tractar la subcategorització dels verbs en català mitjançant tècniques d'aprenentatge automàtic, classificant-los en classes sintàctiques. És a dir, l'objectiu és construir un sistema que sigui capaç d'adquirir automàticament informació sobre els verbs i, en concret, sobre quin tipus de complements poden portar. La informació sobre quins elements subcategoritza cada verb és molt important, ja que determina, en gran part, l'estructura sintàctica de la frase. Per tant, pot resultar molt útil per a moltes tasques de processament del llenguatge natural (PLN).

Actualment, s'han aconseguit desenvolupar mètodes automàtics d'anàlisi morfològica i sintàctica superficial força acurats i eficients. Tanmateix, per poder seguir millorant els resultats que ofereixen les eines de PLN cal tenir accés a informació lèxica, informació sobre el comportament lingüístic de cada paraula (en quines col.locacions s'usa, quins complements demana, etc.). Si fa uns anys el lexicó es considerava un element secundari, tant en tasques de PLN com en lingüística teòrica, actualment pren cada cop un paper més central. Per exemple, hi ha teories lingüístiques afirmen que tot el coneixement lingüístic és coneixement sobre paraules i, cada cop més, les teories lexicalistes actuals tendeixen a prendre coneixement de la gramàtica i a introduir-lo en les entrades lèxiques (per exemple, HPSG (Sag i Pollard, 1987) o LFG (Bresnan, 2001)). A més, un sistema que pretengui analitzar text no restringit, cal que sigui suficientment robust, per la qual cosa ha de disposar d'informació lèxica.

En aquest context, el lexicó és cada cop més important i es fa evident la necessitat de disposar de més informació lèxica. Tanmateix, no és sempre fàcil disposar d'aquest tipus d'informació. Recollir informació lèxica manualment (sobre la subcategorització dels verbs en el cas que tractarem aquí) és un procés lent i costós, el resultat del qual, a més, sovint presenta incoherències i mai no pot ser completament exhaustiu. L'adquisició lèxica és una especialitat de la lingüística computacional (LC) que pretén contrarestar aquestes mancances inferint propietats lèxiques dels mots a partir, entre d'altres, del seu comportament en grans corpus mitjançant tècniques estadístiques i d'aprenentatge automàtic. Els corpus, especialment aquells anotats amb informació morfològica i sintàctica, es veuen com un "dipòsit" de gramàtiques implícites, que es poden explotar mitjançant tècniques automàtiques (Merlo i Stevenson, 2001, pàgina 399).

Com hem dit, la informació sobre subcategorització verbal és clau en les entrades lèxiques i, per tant, si la coneixem, ens permetrà produir anàlisis sintàctiques més acurades. Per exemple, disposar d'informació de subcategorització pot ajudar a un *chunker* (un segmentador de sintagmes) a reconèixer les dependències sintàctiques dins una frase. Evidentment, també serà útil per a tasques més complexes: per exemple, per tal que un analitzador sintàctic pugui decidir si un sintagma preposicional situat després del verb és un complement de règim o un complement circumstancial (*creure en Déu* vs. *comprar en un mercat*).

L'objectiu d'aquest treball és presentar un sistema capaç de classificar verbs automàticament segons el tipus de complements que admeten. Per tal d'aconseguir-ho, definirem trets que ens puguin ajudar a distingir diferent tipus de verbs i que es puguin detectar automàticament en un corpus, amb l'objectiu que després pugem inferir la classe d'un verb a partir del seu comportament en un corpus.

Aquest treball s'estructura de la manera següent:

- El capítol 2 conté una revisió bibliogràfica d'alguns treballs previs sobre adquisició lèxica i sobre alguns aspectes referents a la subcategorització verbal.
- El capítol 3 exposa els experiments i resultats obtinguts fent servir un mètode supervisat. Es pretén distingir automàticament entre verbs transitius i intransitius amb un sistema que, a partir d'un conjunt de verbs prèviament anotats, dedueix regles per a detectar cada classe.
- El capítol 4 presenta els experiments i resultats obtinguts amb un mètode no supervisat, el *clustering*. Aquest mètode no parteix de verbs prèviament anotats, sinó que agrupa els verbs basant-se en les semblances i diferències dels valors dels trets que els descriuen.
- El capítol 5 mostra dues línies d'experiments addicionals que hem fet a partir dels presentats en el capítol 4. D'una banda, a la secció 5.2 mostrem com es pot emprar el nostre sistema de classificació amb un corpus etiquetat automàticament. D'altra banda, la secció 5.3 presenta uns experiments preliminars per poder adquirir automàticament els verbs que exigeixen un complement preposicional.
- Finalment, el capítol 6 presenta les conclusions d'aquest treball de recerca.

Revisió bibliogràfica

2.1 Introducció

En aquest treball cal aplicar coneixements de dos camps diferents:

1. D'una banda, la sintaxi i, més concretament, la subcategorització verbal. En la secció 2.2, revisarem alguns aspectes referents a la sintaxi del sintagma verbal.
2. De l'altra banda, l'adquisició lèxica. En la secció 2.3, farem una revisió dels principals treballs que s'han dut a terme en aquest camp.

2.2 El verb

El nostre objecte d'estudi i de classificació són els verbs. Com hem explicat a la introducció, els verbs condicionen en gran mesura l'estructura de la frase, ja que determinen quants i quin tipus d'arguments hi haurà, així com quines característiques semàntiques tindran. Cada verb presenta un determinat marc de subcategorització (MSC), format pel conjunt de categories sintàctiques amb les quals pot aparèixer. El nostre objectiu serà agrupar els verbs segons el seu comportament sintàctic, cosa que farà més fàcil l'anàlisi de la frase en què apareguin.

El lèxic ha anat adquirit cada cop més importància dins la teoria lingüística, ja que es fa servir per capturar diverses generalitzacions lingüístiques. Molts models sintàctics lexicalistes (LFG, HPSG) tracten la subcategorització i altres propietats sintàctiques en les entrades lèxiques, de manera que el component sintàctic es redueix. Per exemple, en HPSG (Sag i Pollard, 1987), l'entrada lèxica d'un verb transitiu inclou informació sobre el seu subjecte i objecte.

2.2.1 El sintagma verbal dins l'oració

El sintagma verbal (SV) és la realització estructural de la predicació verbal (Rosselló, 2002). Tanmateix, des del punt de vista de la lògica moderna, un predicat s'oposa a tots els seus

arguments (subjecte inclòs): per tant, *donar* té tres arguments: l'agent, el tema i el beneficiari. En una frase activa, l'agent realitza la funció de subjecte; el tema, la funció de complement directe; i el beneficiari, la de complement indirecte. Per a molts, el subjecte és un argument extern, ja que no pertany al sintagma verbal, mentre que els altres arguments són interns.

El verb té dos tipus d'elements que el complementen: els arguments i els adjunts. Els arguments són les entitats referencials que participen en la situació denotada pel predicat. Cada verb determina el nombre d'arguments que necessita, així com alguna de les seves característiques semàntiques i la seva concreció sintàctica en un sintagma determinat. Els adjunts, en canvi, no tenen una relació tan estreta amb el verb.

Per la nostra recerca, ens interessa detectar el comportament sintàctic del verb, conèixer de quins elements pot o ha d'anar acompanyat superficialment. Per això, dins aquest apartat, revisarem alguns tipus de verbs (transitius, intransitius, inacusatius, règim verbal) i n'establirem les característiques principals.

2.2.1.1 Els verbs transitius

El concepte de transitivitat ha rebut moltes definicions. Per exemple, per Fabra (1956), els verbs transitius són els “que denoten una acció que recau sobre una persona o cosa expressada per un complement acusatiu”. Segons Lyons (1968), “un verb transitiu és aquell que expressa una acció que transcendeix d'un agent a un pacient”. D'altres autors, com Hernanz i Brucart (1987), rebutgen aquesta mena de definicions nocionals, ja que podem trobar molts casos de verbs transitius l'objecte dels quals no “rep” cap mena d'acció:

- (1) La Maria té por

Hernanz distingeix entre transitivitat directa (presència del complement directe), transitivitat indirecta (és a dir, amb mediació de preposició, cosa que inclouria els verbs de règim verbal) i doble transitivitat (quan hi ha dos elements subcategoritzats).

Des del punt de vista sintàctic, els verbs transitius tenen la propietat d'exigir un sintagma objecte. Cano (1987) caracteritza la transitivitat amb les següents propietats:

1. Indeterminació semàntica del verb quan li manca l'objecte
2. Cohesió verb-objecte, manifestada en la integració de significat i la unitat rítmica
3. Possibilitat de pronominalitzar l'objecte

4. La pregunta per l'objecte és *què* o *a qui*
5. Possibilitat d'admetre la passiva

Tanmateix, hi ha verbs que no compleixen alguna d'aquestes característiques. Per exemple, hi ha verbs que no admeten la passiva o verbs que poden aparèixer sense objecte. Segons Hernanz i Brucart (1987), el criteri de la pronominalització, tot i que no sempre és decisiu, és probablement el més fiable.

Dir si un verb és transitiu no és una tasca fàcil, ja que ens podem trobar amb diversos problemes (Rosselló, 2002):

1. Certs sintagmes que han patit fenòmens d'incorporació lèxica, com *fer badalls* o *fer feina*, no presenten cap expressió referencial definida i, de fet, són sinònims de predicatius intransitius com *badallar* o *treballar*. Tanmateix, des del punt de vista sintàctic, presenten un patró transitiu, ja que són cliticitzables amb el pronom *en* (*En fa molta, de feina*).
2. Certs verbs, com *beure* o *menjar*, poden aparèixer sense cap marca formal de transitivitat, però necessiten un tema des del punt de vista semàntic (*Ni dormo, ni menjo,estic malalta*).
3. Els predicats que seleccionen una expressió nominal de mesura (*Aquesta pel·lícula dura dues hores*) o els verbs implicatius (*Això significa la derrota definitiva*) no accepten la passiva o la cliticització acusativa.
4. Alguns verbs intransitius poden realitzar-se amb un acusatiu intern (*Encara vivim la vida misèrrima dels pobres*).
5. Els verbs inacusatius també són problemàtics, ja que el seu únic argument té un comportament més d'objecte que de subjecte (*Falten tres bitllets*). En parlarem a la següent secció.

2.2.1.2 Els verbs intransitius i inacusatius

Els verbs intransitius no porten complement directe i només tenen el subjecte com a argument. Alcina i Bleuca (1975) els divideixen en les següents classes:

- Verbs existencials com *existir*, *morir*, *viure*. Alguns d'aquests verbs permeten un complement tautològic, amb la qual cosa es construeixen com si fossin transitius:

(2) El meu avi va morir una mort piatosa

- Verbs de moviment com *caminar, pujar, entrar, anar, saltar, viatjar, tornar*, etc. Alguns admeten un complement directe:

(3) La nena va baixar l'escala sola

- Verbs d'acció com *riure, cridar, suar, tremolar, bordar, tossir*. Alguns poden aparèixer amb complement directe:

(4) El malalt va tossir sang

- Verbs pseudoimpersonals com *cabre, agradar, faltar, importar, molestar, sobrar*. El subjecte sol ser inanimat i apareixen amb un complement indirecte:

(5) Al Joan no li agraden els ordinadors

La majoria de marcs teòrics en sintaxi distingeixen dos tipus de verbs intransitius: els inacusatius (o ergatius) i els intransitius purs (o inergatius)(Levin i Rappaport, 1995).

Els inacusatius són els verbs l'argument del qual presenta un comportament més d'objecte que de subjecte (Rosselló, 2002). El seu paper temàtic respon més al de tema que al d'agent, cosa que no passa amb els intransitius purs. Els intransitius purs solen designar activitats que depenen de la voluntat d'un agent (*cridar, córrer, riure*, etc.). Els inacusatius, en canvi, designen assoliments o bé estats no agentius. Els que designen assoliments es poden dividir en moviment de direcció inherent (*venir, arribar*) i verbs presentacionals (*aparèixer*). Els inacusatius, com els intransitius purs, només tenen un argument. Tamateix, en el cas dels inacusatius, aquest argument s'interpreta com l'objecte lògic o semàntic, tal i com ocorre amb l'objecte dels verbs transitius.

Les principals diferències sintàctiques entre intransitius purs i inacusatius són les següents:

1. El subjecte dels inacusatius es pot postposar amb molta més facilitat que el dels intransitius. De fet, la posició potsverbal és la "típica" dels subjectes dels verbs inacusatius. Aquest fet dificultarà molt la distinció automàtica entre transitius i inacusatius (vegeu el capítol 3).
2. El subjecte postverbal dels inacusatius pot aparèixer sense determinant.

(6) a. Falten nens

b. * Riu gent

3. Els subjectes indefinites dels inacusatius poden substituir-se pel pronom *en*¹.

- (7) a. N'han arribat dos
b. * En baden quatre

2.2.1.3 Verbs de règim verbal

El complement preposicional és un complement subcategoritzat pel verb que va introduït per preposició, propietat que comparteix amb el complement indirecte. Tanmateix, a diferència del complement indirecte, hi trobem moltes altres preposicions apart de *a* (*en, de, amb, per*) i no suposa l'existència d'un complement directe.

A diferència del cas dels complements circumstancials, en els complements preposicionals la preposició és semànticament buida. Per tant, mentre que un complement circumstancial introduït per *en* indicarà un locatiu i la preposició podrà ser substituïda per una altra de valor semblant, això no serà possible si es tracta d'un complement preposicional (Hernanz i Brucart, 1987).

- (8) a. En casa hace frío
b. Dentro de casa hace frío
c. Confías demasiado en la vecina
d. * Confías demasiado dentro de la vecina

Segons Cano (1999), els verbs pronominals (els verbs incrementats per la partícula *se*) mostren una notable tendència a exigir sintagmes preposicionals. En alguns casos, aquests sintagmes coincideixen amb el complement que apareix amb el verb en forma no pronominal seguint el complement directe: *dedicar su vida a algo* → *dedicarse a algo*. El clíctic seria una marca d'aquest objecte. En altres casos, el complement preposicional del verb pronominal no apareix en la forma no pronominal: *olvidar algo/olvidarse de algo*.

2.2.2 Alternances

En aquest apartat, revisarem les principals alternances que trobem en els verbs en català. Es dona una alternança quan un verb pot aparèixer en diversos formats oracionals. Per tant, un

¹La distinció resulta poc clara en alguns casos, ja que certs verbs que semblen clarament intransitius purs també admeten aquest tipus de substitució, especialment si hi ha locatius a la frase.

verb que participi en una alternança no tindrà un sol patró de comportament en el corpus, cosa que en pot dificultar la classificació automàtica.

Les alternances poden ser de valència o de diàtesi. Un verb presenta una alternança de valència quan pot ocórrer amb un nombre diferent d'arguments. L'exemple (9) és gramatical tant amb complement directe com sense.

(9) En Pere menja ((els) entrepans)

La diàtesi és la relació entre funcions semàntiques i funcions sintàctiques. Per tant, una alternança de diàtesi es produeix quan aquesta relació presenta més d'un model amb un mateix verb.

(10) a. La mestra ha obert la finestra

b. La finestra s'ha obert

Levin (1993) proposa una ampla classificació dels verbs per a l'anglès segons les seves alternances de diàtesi, classificació que, com veurem, s'ha emprat àmpliament pels treballs d'adquisició lèxica. El treball de Levin parteix del pressupòsit que els verbs amb un comportament semblant (amb les mateixes alternances) tenen també un significat semblant.

Els principals tipus d'alternances per al català són (Rosselló, 2002):

1. Complement directe / \emptyset : Un verb que semànticament requereix dos arguments pot presentar un alternant intransitiu, és a dir, amb el complement directe absent. Per exemple:

(11) El nen ja ha menjat

Se sol considerar que aquests verbs nocionalment continuen sent transitius, ja que sembla clar que, en l'exemple anterior, se sobreentén que hi ha alguna cosa que ha estat menjada. En principi, poden patir aquesta alternança tots els verbs transitius que no denoten estats (**En Joan tindrà*).

2. Complement indirecte / \emptyset : Els verbs transitius de transferència seleccionen semànticament una meta que es pot realitzar superficialment o no:

(12) He donat tots els meus llibres (a la biblioteca municipal)

Els verbs psicològics intransitius (*agradar, doldre*) i els de contacte (*pegar*) també seleccionen un complement datiu que es pot ometre.

3. Complement directe / complement indirecte: Certs verbs psicològics apareixen tant complementats amb un acusatiu com amb un datiu:

- (13) a. Això no li preocupa
b. Això no la preocupa

4. Complement directe / complement de règim: Aquest és un cas d'alternança de diàtesi: el complement del verb es pot materialitzar com a complement directe o com a complement de règim. Per exemple:

- (14) a. Recórrer la sentència
b. Recórrer contra la sentència

En alguns casos, les dues alternances no són sinònimes:

- (15) a. Cridar algú
b. Cridar a algú

Hernanz i Brucart (1987) també assenyalen un grup de verbs que presenten la següent alternança (n'hem parlat també a 2.2.1.3):

- (16) a. lamentar SN / lamentarse de SN
b. olvidar SN / olvidarse de SN
c. confesar SN / confesarse de SN
d. aprovechar SN / aprovecharse de SN

En aquests cas l'aparició de la preposició es troba supeditada a la presència de la marca *se*. La combinació *V amb clític concordat + acusatiu* i la combinació *V + complement de règim* són impossibles (Rosselló, 2002):

- (17) a. *admirar* + acusatiu
b. *admirar-se de*
c. * *admirar-se* + acusatiu
d. * *admirar de*

5. Complement de règim: Es tracta d'una alternança en què no canvia ni la valència ni la diàtesi, sinó que el sintagma preposicional que fa de complement pot anar introduït

per més d'una preposició: *parlar de/sobre, protegir-se de/contra*. En alguns casos, l'alternança implica variació de significat:

(18) a. El gripau es va convertir en un príncep

b. S'ha convertit al cristianisme

6. Complement de règim / \emptyset : Verbs que regeixen un complement preposicional amb una alternança intransitiva: *jugar (jugar a), presumir (presumir de)*.

7. Complement locatiu / \emptyset : Alternança de diàtesi en què el complement directe i l'oblic correspon al tema i locatiu en una alternança i, en l'altra, a la inversa.

(19) a. En Joan va carregar les taronges a la furgoneta,

b. En Joan va carregar la furgoneta de taronges

8. Alternança causativa-anticausativa.

L'alternança causativa-anticausativa és una alternança diatètica entre un ús transitiu i un ús intransitiu.

(20) a. El nen ha trencat el vidre

b. El vidre s'ha trencat

El subjecte de l'estructura intransitiva té el mateix paper temàtic que l'objecte de l'estructura transitiva; concretament, corresponen a l'entitat afectada, al tema (Vázquez, 1997). En la transitiva, a més, el subjecte expressa la causa de l'event. En català, aquesta alternança té efectes en la forma del verb, que sol aparèixer amb el clític *se*, tot i que hi ha verbs causatius que no admeten aquesta forma pronominal (*millorar, empitjorar, augmentar, minvar, bullir, coure, envellir*).

Per tant, en català, els verbs poden aparèixer en moltes alternances diverses. Aquest fet dificultarà l'adquisició de marcs de subcategorització, ja que pot succeir que els verbs que pateixin alternances tinguin un comportament poc uniforme respecte a les classes que haguem definit. Per exemple, si volem distingir exclusivament entre verbs transitius i intransitius, és evident que els verbs que participen en l'alternança "complement directe / \emptyset " seran problemàtics, ja que són verbs transitius que, de vegades, però no sempre, es comporten com a intransitius (vegeu el capítol 3). Tanmateix, també ens podem aprofitar del fet que un grup de verbs aparegui molt regularment en els diversos formats oracionals d'una alternança per intentar aprendre automàticament quins verbs participen en aquesta alternança (vegeu el capítol 4).

2.2.3 La partícula *se*

L'anàlisi de la partícula *se* és problemàtica, ja que presenta molts matisos, apareix amb tipus de verbs diferents i provoca diferents conseqüències en l'estructura sintàctica de la frase. A més, apareix amb molta freqüència. Segons Cano (1987), una quarta part de les formes verbals que apareixen en un text en espanyol van incrementades per *se*, de les quals només un petit percentatge són pròpiament reflexives. Segons aquest autor, la major part dels verbs incrementats per *se* indiquen un procés desenvolupat en el subjecte. Segons Bartra (2002) podem distingir diferents tipus de *se*:

1. *Se* reflexiu. La partícula *se* apareix quan l'objete directe o indirecte és idèntic al subjecte, amb la qual cosa sembla eliminar el tema.

(21) a. En Joan es pentina

b. En Joan em pentina

Dins d'aquest grup, també trobem el *se* recíproc, que apareix en frases en què dos o més subjectes realitzen una acció que recau sobre l'altre o altres. El pronom *se* pot alternar amb pronoms reflexius d'altres persones gramaticals: *em*, *ens*, etc.

(22) a. En Joan i la Maria es detesten cordialment

b. En Joan i la Maria em detesten cordialment

2. *Se* datiu possessiu. Apareix en frases on hi ha un nom de possessió inalienable i alterna amb datius no reflexius.

(23) a. En Jaume s'ha afaitat la barba

b. En Jaume m'ha afaitat la barba

3. *Se* pronom aspectual. Segons Bartra (2002), contribueix a establir la telicitat del predicat i també pot alternar amb altres pronoms.

(24) a. Ja s'ha menjat el bistec

b. Ja m'he menjat el bistec

4. *Se* inherent. Apareix en verbs pronominals en què el pronom reflexiu no té antecedent i alterna amb altres pronoms.

(25) a. No es penedeix de res

b. No em penedeixo de res

5. *Se* en construccions pronominals de subjecte inespecífic. La presència del clític elimina l'aparició de l'agent i no es pot intercanviar per pronoms d'altres persones. Són les oracions també denominades passives pronominals o reflexes.

- (26) a. S'han iniciat les negociacions de pau
 b. Si es neix ric, tot és més fàcil

2.2.4 Algunes propostes de classificació

Els verbs es poden agrupar o classificar seguint molts criteris diferents: morfològics, sintàctics, semàntics o pragmàtics, diversitat donada per les característiques pròpies del verb (Lorente, 1996). Per als nostres objectius, és evident que els criteris sintàctics seran els que més ens ajudaran. Veiem algunes propostes:

- Classificació de Fabra (1956). Fabra classifica els predicats en quatre tipus:
 - Tipus I: El nen plora
 - Tipus II: Aquest regle és curt
 - Tipus III: Plou. És clar.
 - Tipus IV: Ha arribat un parent meu

Fabra barreja criteris sintàctics i semàntics en la seva classificació. Per exemple, *Plou* (tipus III) també té característiques comunes amb les frases del tipus I (el verb no és copulatiu); i *És clar* (tipus III) també és copulativa (com les frases del tipus II). En conseqüència, no és una classificació adequada per dur a terme experiments sobre adquisició de classes sintàctiques de verbs.

- Classificació de Lorente (1996). Lorente (1996) presenta una classificació basada en la poliadicitat (Bresnan, 1982), el nombre d'arguments que pot portar una forma verbal.
 - Anàdics (sense arguments)
 - * *Ploure*
 - Monàdics (un argument)
 - * Amb argument extern: *viatjar*
 - * Amb argument intern: *arribar*
 - Diàdics (dos arguments)
 - * Amb argument extern: *menjar*

- * Amb argument intern: *agradar*
- Triàdics (tres arguments)
 - * Amb argument extern: *donar*
 - * Amb argument intern: *rebre*

El fet d'introduir els verbs anàdics permet donar compte dels verbs meteorològics i d'altres que presenten la llista de subcategorització del subjecte buida. Aquesta classificació està basada en criteris molt més clars i definits que la de Fabra (1956). Tanmateix, un punt conflictiu per a l'aplicació d'aquesta classificació és la distinció entre adjunts i arguments, especialment en el cas del complement indirecte. Seria un objectiu molt ambiciós pretendre adquirir totes aquestes classes automàticament, a partir dels recursos dels que disposem (vegeu la secció 3.2). Caldria, com a mínim, un corpus analitzat sintàcticament per detectar els verbs que subcategoritzen més d'un argument. A més, aquesta classificació no inclou distincions que sí que es podrien adquirir automàticament amb més facilitat: verbs que subcategoritzen completives, verbs que regeixen complements preposicionals, etc.

- Classificació de Rosselló (2002). Aquesta autora presenta una classificació dels “principals patrons categorials en què es realitzen els arguments en el sintagma verbal de l'oració simple” (Rosselló, 2002, pàgina 1928). Dins de cada apartat, trobem dos patrons: el primer té argument extern i el segon no (és a dir, es tracta d'oracions inacusatives).
 1. (sv V)
 - SN (sv V). Oracions amb verb intransitiu pur: *El Pere dorm.*
 - __ (sv V). Oracions amb verb atmosfèric: *Plou.*
 2. (sv V SN)
 - SN (sv V SN). Oracions transitives més simples: *El veí estenia roba.*
 - __ (sv V SN). Oracions inacusatives més simples: *Cal gent.*
 3. (sv V SN SP)
 - SN (sv V SN SP). Oracions transitives amb verbs de transferència: *La Carme regalava llibres als alumnes.*
 - __ (sv V SN SP). Oracions inacusatives biargumentals: *Han tocat dos llibres a cada nen.*

4. (sv V SP)

– SN (sv V SP). Oracions amb complement de règim verbal: *En Quim confia en la sort.*

– __ (sv V SP). Patró inexistent.

5. (sv V SP SP)

– SN (sv V SP SP). Oracions amb verbs com *tractar (d'x amb y)*, *discutir (de/sobre x amb y)* o *col.laborar (amb x en y)*: *En Pere col.labora amb en Jordi en un projecte.*

– __ (sv V SP SP). Patró inexistent.

6. (sv V SN SP SP)

– SN (sv V SN SP SP). Oracions amb verbs de transacció comercial: *El ministre va comprar dues finques a aquells gàngster per mil milions de pessetes.*

– __ (sv V SN SP SP). Oracions inacusatives de verb de direcció inherent: *En passar els nens de l'escola a l'institut, ja és una altra cosa.*

Aquesta classificació parteix de criteris molt coherents i clars i és molt exhaustiva. Aquesta exhaustivitat fa que sigui una classificació molt difícil d'adquirir automàticament i que queda molt per sobre dels nostres objectius. Com en el cas de la classificació anterior, també caldria un corpus amb informació sintàctica per adquirir totes les classes proposades.

2.2.5 Conclusions

Aquesta breu revisió d'alguns fenòmens relacionats amb el sintagma verbal ens ha permès de caracteritzar les classes que voldrem adquirir automàticament: verbs transitius, intransitius, verbs que participen en certes alternances, etc. Les característiques pròpies de cada classe ens serviran per establir els trets que emprarem pels experiments d'adquisició de classes verbals (vegeu 3.3).

Aquest apartat també ens permet preveure alguns dels aspectes que resultaran problemàtics: com distingir els verbs inacusatius dels transitius (ja que tots dos porten un argument darrere el verb), quins efectes té la presència de la partícula *se* sobre els verbs transitius o com cal tractar els verbs que participen d'alguna alternança i que, per tant, apareixeran en el corpus en diferents contextos lingüístics.

Hem vist també que les classificacions revisades no s'adapten als nostres objectius. Les classificacions que es basen en criteris sintàctics clars (Lorente, 1996) (Rosselló, 2002) són

massa ambicioses, ja que distingeixen massa classes diferents. Tenint en compte el recurs dels quals disposem, en els capítols 3 i 4 proposarem unes classificacions més generals, que siguin possibles d'adquirir, i a partir de les quals es puguin adquirir subclasses més específiques.

2.3 L'adquisició lèxica

L'objectiu de l'adquisició lèxica és desenvolupar tècniques estadístiques que permetin adquirir propietats de les paraules inferint-les del seu comportament en grans corpus textuais o diccionaris. El coneixement extret a partir de tècniques d'adquisició lèxica pot resultar interessant principalment per dues raons:

1. No hi solen haver recursos que contemplin aquest tipus d'informació. Si bé existeixen diccionaris electrònics, aquests són incomplets, presenten incoherències i, normalment, no ofereixen informació sobre freqüències.
2. El fet d'adquirir informació de corpus textuais ofereix majors garanties empíriques, ja que les dades que s'extreuen no depenen en tan gran mesura dels pressupòsits o opinions de qui analitza.

Els mètodes d'adquisició lèxica s'han aplicat per resoldre diverses tasques (per exemple, desambiguació de sentits (Pereira, Tishby i Lee, 1993) o desambiguació de l'adjunció del sintagma preposicional (Hindle, 1993)) i per tractar diverses categories o fenòmens lingüístics: per exemple, col.locacions (Dunning, 1993), (Justeson, 1995), adjectius (Boleda, 2003), classes semàntiques (Zernik, 1989), preferències de selecció (Resnik, 1993), (Ribas, 1995), alternances de diàtesi (Merlo i Stevenson, 2001), (Lapata, 1999), (Lapata, 2000) marcs de subcategorització (Brent, 1991), (Brent, 1993), (Ushioda et al., 1993), (Briscoe i Carroll, 1997), (Manning, 1993), (Carroll, 1998b), etc. La nostra revisió se centrarà en els treballs sobre adquisició verbal (per una revisió més completa vegeu Matsumoto (2002)).

2.3.1 Treballs previs d'adquisició de la subcategorització dels verbs

En el camp de l'adquisició lèxica, el verb és la categoria morfològica que més atenció ha rebut (Brent, 1991), (Brent, 1993), (Manning, 1993), (Ushioda et al., 1993), (Briscoe i Carroll, 1997), (Eckle, 1996), (Rooth et al., 1999). Això és així ja que els verbs són la principal font d'informació en una frase, especialment pel que fa a l'estructura. El fet de disposar d'informació de subcategorització verbal ens pot ajudar a fer una millor anàlisi sintàctica

(*parsing*). De fet, segons Carroll (1998a), més del 50% dels errors de *parsing* es produeixen a causa de manca d'informació sobre la subcategorització dels verbs. Sense aquesta mena d'informació, distingir arguments d'adjunts o resoldre la majoria d'ambigüitats d'adjunció de sintagmes es fa molt difícil.

2.3.2 Els primers treballs: compilació de marcs de subcategorització

Els primers sistemes capaços d'extreure informació de subcategorització de forma automàtica a partir de corpus van començar a aparèixer per primer cop fa més d'una dècada (Brent, 1991), (Brent, 1993).

Molts dels primers treballs sorgits en aquest camp no s'orientaven cap a una classificació automàtica dels verbs segons la seva subcategorització, sinó que pretenien compilar els possibles marcs de subcategorització de cada verb, per tal de construir diccionaris (en són exemples: Ushioda et al. (1993), Manning (1993), Carroll (1998b)). Un dels treballs pioners és del de Brent (1991) que pretén extreure un diccionari de marcs de subcategorització (MSC).

Brent va proposar un mètode per adquirir cinc MSC diferents (objecte directe, objecte directe + oració de relatiu, objecte directe + infinitiu, oració de relatiu, infinitiu); és a dir, MSC que només incloguessin sintagmes nominals, oracions de relatiu i infinitius. Cal destacar que Brent (com Manning (1993) o Briscoe i Carroll (1997)) entén MSC com cada combinació diferent d'elements que pot complementar un verb.

Brent (1991) pretenia explotar informació no ambigua de corpus no anotat i va definir correlats lèxics (*lexical cues*) segurs i útils per detectar verbs i els seus MSC. Brent identificava els potencials verbs cercant mots que apareguessin en el corpus amb la flexió *-ing* i sense aquesta flexió. Un cop trobat, si el mot no anava precedit de determinant o d'una preposició diferent de *to*, el considerava com a verb. En cas de dubte, ignorava aquesta ocurrència. Evidentment, aquesta és una aproximació simplista que provoca errors i un nivell de cobertura molt baix, ja que, fent servir aquests mètodes, molts verbs no es poden detectar.

Un cop trobat un verb, es buscava si hi havia els patrons que corresponien als MSC definits. Per exemple, si, a la dreta d'una paraula detectada com a verb es trobava el correlat [*to V*], aquesta seqüència de paraules es comptabilitzava com a complement d'infinitiu (com a, per exemple, *I hope to attend*). Tot i que Brent fa servir correlats molt segurs, aquest sistema no està exempt de soroll. Per exemple, el verb anglès *refer* es classificaria erròniament com a verb que porta un complement d'infinitiu en la frase *I referred to changes made under the military occupation*. Brent va intentar resoldre aquest problema aplicant filtres estadístics.

En general, aquesta sistema tenia una bona precisió, però una cobertura baixa, ja que no disposava d'un corpus anotat i només tractava els casos no ambigus. A més, depenia dels correlats lèxics per detectar MSC, amb la qual cosa mai no hauria pogut detectar els MSC que no tenen un correlat lèxic clar. Per exemple, aquest sistema no pot detectar els verbs que subcategoritzen la preposició *in* (*They assist the police in the investigation*), ja que la majoria d'ocurrències d'aquesta preposició després d'un verb introdueix adjunts (*He built a house in the woods*).

Com que només va tractar casos no ambigus, Brent (1991) no ofereix informació sobre freqüència, informació molt valuosa per a qualsevol lexicó computacional. Per això, estudis posteriors han intentat resoldre aquests problemes i treure dades de tots els exemples de les dades d'entrenament (i no només dels no ambigus), cosa que fa necessari un corpus anotat morfològicament o sintàcticament.

Ushioda et al. (1993) i Manning (1993) persegueixen un objectiu semblant al de Brent (1991), però proposen un mètode més sofisticat i analitzen sintàcticament el text, encara que de forma parcial, a través d'un *parser* d'estats finits.

Ushioda et al. (1993) fa servir un corpus etiquetat morfològicament i un *parser* d'estats finits per reconèixer sis tipus de classes de subcategorització (les mateixes que els sistemes de Brent, a més de "sintagma nominal + sintagma nominal"). S'extreu el verb, es fa l'anàlisi sintàctica parcial mitjançant un *chunker* (un segmentador de sintagmes) i s'obté el possible MSC. Els resultats presenten errors i casos problemàtics, fruit dels problemes amb què ja es va trobar Brent (SN compostos, dificultat per distingir entre complements i adjunts, etc.). Ushioda et al. (1993) fa servir un mètode estadístic adicional que aprèn dels errors i cada cop ofereix resultats més acurats. Arriba a una precisió del 83% en classificar 1565 tokens de 33 verbs. És un dels primers treballs que ofereix informació sobre freqüències.

Manning (1993) proposa un sistema similar al d'Ushioda et al. (1993), però més ambiciós, ja que pot reconèixer 19 MSC. Alguns dels MSC que pot reconèixer són: verbs transitius, intransitius, ditransitius, verbs complementats amb completiva, verbs complementats amb infinitiu, etc. Manning fa servir un *tagger* estocàstic i un *parser* d'estats finits. Quan el *parser* detecta un verb, analitza els complements que el segueixen fins que troba un element que marca el final del MSC (un signe de puntuació o una conjunció subordinada). Posteriorment, fa servir un procés de filtratge que li permet eliminar alguns errors (el *parser* sempre tracta els adjunts com a complements). El sistema de Manning adquireix, per a 3104 verbs, 4900 MSC, alguns dels quals no es troben en els pocs diccionaris de subcategorització que existeixen. La precisió és del 90% i la cobertura del 43% (Manning calcula aquestes mesures comparant els MSC que

el seu sistema adquireix i els MSC que figuren en un diccionari).

Briscoe i Carroll (1997) persegueixen el mateix objectiu que els treballs que acabem de presentar, però construeixen un sistema molt més ambiciós, ja que reconeix fins 160 tipus de subcategorització, principalment extrets de dos diccionaris (ANLT i COMPLEX) o documentats en corpus. El sistema consta d'un *tagger*, un lematitzador, un *parser* probabilístic, un sistema per extreure models de subcategorització, un classificador i un avaluador. Aquest sistema arriba a un nivell d'encert del 81%, resultat certament impressionant si considerem el nombre de tipus que hi ha i, a més, ofereix informació de freqüència de cada MSC per a cada verb.

En resum, els primers treballs minimitzaven el soroll, amb la qual cosa la cobertura es reduïa (Brent, 1991), (Brent, 1993). Els treballs següents van maximitzar la cobertura, amb la qual cosa la precisió empitjorava (Ushioda et al., 1993), (Manning, 1993), i treballs més recents han intentat maximitzar tant la cobertura com la precisió.

2.3.3 Classificació automàtica

Més recentment s'han disenyat sistemes que empren sistemes d'aprenentatge automàtic (*machine learning*) per classificar verbs en diferents classes, segons les seves característiques sintàctiques o semàntiques. Els algorismes d'aprenentatge automàtic es poden subdividir en dos grans grups: els supervisats i els no supervisats.

Els mètodes supervisats parteixen d'un conjunt d'objectes (que, en el nostre cas, seran lemes verbals) prèviament classificats o anotats. A partir de les dades de què es disposa sobre cada objecte (entre les quals hi ha la informació sobre la classe a la qual pertany), els mètodes supervisats intenten extreure generalitzacions sobre el comportament de cada classe. Això els permetrà, posteriorment, de predir la classe d'un objecte nou, que no pertany al grup d'objectes prèviament anotats. En canvi, en els mètodes no supervisats, els objectes no estan prèviament classificats i l'algoritme simplement els agrupa segons les dades que els descriuen. Per exemple, un dels mètodes no supervisats més utilitzats, el *clustering*, classifica els objectes en un cert nombre de grups (*clusters*), de manera que cada grup contingui objectes que s'assemblin al més possible entre ells i que siguin diferents dels objectes dels altres clusters.

Els mètodes no supervisats, per tant, no depenen d'una classificació prèvia, sinó que, per fer els *clusters*, empren només la informació que tenen de cada objecte. En aquest sentit, podríem considerar que el *clustering* és un mètode més "objectiu", ja que no està lligat a cap classificació. Tanmateix, aquest fet té el desavantatge que la classificació proposada per

l'algoritme no té perquè coincidir amb la divisió que es voldria aconseguir, ja que l'algoritme potser fa unes generalitzacions sobre les dades que no tenen cap mena de rellevància per a la classificació que es pretenia fer. Evidentment, si els trets són prou distintius i rellevants, això no hauria d'ocórrer.

Els mètodes supervisats, en canvi, tenen l'avantatge que les regles que l'algoritme deduirà segur que tindran relació amb les classes que s'han definit prèviament. El gran problema d'aquesta aproximació és que cal disposar d'una classificació fiable, coherent i completa, cosa no gens fàcil.

El treball de Merlo i Stevenson (2001) és un dels que més s'acosta als objectius d'aquest treball, ja que fan servir una aproximació semblant a la nostra per classificar verbs automàticament en tres classes de verbs opcionalment transitius: *inergatius*, *inacusatius* i *object-drop* (és a dir, els que presenten l'alternança complement directe / \emptyset).

Inergatius	
The horse	raced past the barn
AGENT	
The jockey	raced the horse past the barn
AGENT	AGENT
Inacusatius	
The butter	melted in the pan
TEMA	
The cook	melted the butter in the pan
AGENT	TEMA
Object-drop	
The boy	played
AGENT	
The boy	played soccer
AGENT	TEMA

Figura 2.1: Exemples de verbs *inergatiu*, *inacusatiu* i *object-drop* (Merlo i Stevenson, 2001, pàgina 374)

Aquesta classificació és interessant perquè les tres classes tenen marcs sintàctics idèntics (mostren una alternança entre un marc transitiu i un marc intransitiu), però assignen diferents papers temàtics a cada posició (vegeu-ne exemples a la figura 2.1). Saber a quina classe pertany cada verb pot ser útil per a diferents tasques, com pot ser la traducció automàtica, ja que, per exemple, les oracions *inergatives* transitives no són gramaticals en moltes llengües, entre d'altres el català.

En aquest treball, les autores estableixen trets probabilístics que poden ser útils per distingir les tres classes i que tenen correlats lingüístics que es poden extreure automàticament d'un corpus textual suficientment gran. Els cinc trets que aquestes autores estableixen són els

següents:

1. Transitivity: Les autores consideren que un verb té el tret de transitivity si va seguit d'un determinant, adjectiu, nom, pronom o número. En cas contrari, no tindrà aquest tret. Les autores preveuen que els *object-drop* són la classe que tindrà un valor més alt per a aquest tret, ja que són els que presenten usos transitius de manera més no-marcada i amb més freqüència, seguits dels inacusatius i, en últim lloc, dels inergatius (que són rars en construccions transitives).
2. Causativity: Per comptar la causativity, les autores van comptar el grau de solapament dels noms en posicions de subjecte i objecte. Els verbs inergatius i inacusatius tenen usos causatius (el paper temàtic del subjecte de l'ús intransitiu del verb és idèntic al paper temàtic del objecte en l'ús transitiu). En canvi, els verbs *object-drop* no presenten aquest comportament (el paper temàtic del subjecte és el mateix en els usos transitius i intransitius). Aquest tret hauria de servir per distingir els *object-drop* (que no tenen usos causatius) dels altres dos tipus. Tanmateix, com que els inergatius són rars en construccions transitives, les autores preveuen que presentaran un valor baix per aquest tret. Així, la causativity distingirà els inacusatius dels inergatius i *object-drop*.
3. Animacy: El subjecte dels inergatius i dels *object-drop* és agent tant si el verb es fa servir en construccions transitives com en intransitives. En canvi, el subjecte dels inacusatius és tema en construccions intransitives i és agent només en les transitives. Com que els agents solen ser entitats animades, cal suposar que aquest tret pot ajudar a distingir els inacusatius dels inergatius i *object-drop*. Tanmateix, per determinar si una ocurrència d'un verb té aquest tret s'hauria d'anotar manualment o caldria emprar eines on-line, com el WordNet. Per evitar recórrer a recursos externs, les autores proposen que aquest tret tingui un correlat que sí que es pot extreure automàticament de corpus. Les autores assumeixen que els pronoms *I, we, you, she, he i they* fan referència a entitats animades en la major part dels casos. Per tant, consideraran els pronoms com a indicadors de la presència del tret +animat.
4. Ús de la veu passiva. Es comptabilitza el nombre de verbs etiquetats com a participis precedits de l'auxiliar *to be*.
5. Etiqueta morfològica. Es comptabilitza la presència de l'etiqueta morfològica de participi passat.

Aquests últims dos trets contribuiran a marcar si un verb pot tenir un ús transitiu o no; és a

dir, ajuden a reforçar les dades que ja ofereix el tret Transitivity a partir de correlats lingüístics superficials que es poden detectar de manera molt segura automàticament.

Per fer l'estudi, les autores van escollir 20 verbs de cada classe i van comptar la presència de cada tret per a cada verb. Les dades per a cada tret van resultar "roughly as expected" (Stevenson i Merlo, 2001, pàgina 385). El tret Causativity no va donar els resultats que s'esperaven probablement perquè, comptant el solapament de noms en posició d'objecte i subjecte, es poden estar comptant altres fenòmens que no tenen res a veure amb la causativitat, com, per exemple, la reciprocitat.

(27) a. John called his mother

b. His mother called John

Les autores van fer experiments amb diversos mètodes d'aprenentatge automàtic supervisat que fan classificacions automàtiques. Finalment, es van decidir pel sistema d'aprenentatge automàtic C5.0 (Quinlan, 1992), sistema que genera arbres de decisió i indueix regles a partir de les dades. Per als experiments, es van fer servir verbs diferents per a l'entrenament i l'avaluació. La tasca tenia una *baseline* de 33,3% (probabilitat d'encert si es classifica en les tres classes a l'atzar) i, tot i que el màxim d'encert teòric era del 100%, les autores van fer un experiment d'anotació manual humana i l'acord entre especialistes va ser tan sols del 87%.

El seu sistema de classificació automàtica classifica correctament els verbs en un 69,8%, és a dir tan sols un 17% per sota del màxim calculat per les autores. Un dels punts febles del sistema és la distinció entre *inacusatiu* i *object-drop*.

2.3.4 Classificació semàntica

Hi ha una línia diferent de treballs que pretenen classificar els verbs en classes semàntiques. Molts treballs prenen com a punt de partida el treball de Levin (1993), que va classificar 3024 verbs anglesos manualment segons els arguments semàntics que assignen a posicions sintàctiques i segons les seves alternances de diàtesi. La idea de Levin és que "the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning" (Levin, 1993, pàgina 1).

Lapata (1999) proposa un sistema d'adquisició d'alternances verbals. Concretament, se centra en l'alternança de datiu i benefactor, que es manifesta en tres MSC en anglès:

(28) a. Leave her a note

- b. Leave a note for her
- c. John offers shares to his employees

Fent servir una eina anomenada Gsearch (Keller et al., 1999), va extreure patrons que permetessin identificar aquestes alternances; és a dir "V SNI SN2", "V SNI to SN2" i "V SNI for SN2". Com a sortida, va obtenir un gran nombre de frases incorrectes, ja que el *parser* havia analitzat com a casos de doble objecte sintagmes nominals que contenen compostos, oracions de relatiu, aposicions i adjunts. Per evitar aquests errors, Lapata (1999) va postprocessar i filtrar les dades (va detectar noms compostos, possessius, etc).

Lapata i Brew (1999) pretenen tractar aquells verbs que pertanyen a més d'una classe de la classificació de Levin i empen models estadístics per desambiguar la classe. Aquest treball explora fins a quin punt la informació sintàctica es pot fer servir per desambiguar la classe semàntica dels verbs. Per a 306 verbs que mostren un comportament sintàctic diferent segons la classe semàntica a què pertanyen, aconseguen una precisió del 91,8%. En els 31 verbs que mostren el mateix MSC, però que, de fet, poden pertànyer a diferents classes semàntiques, arriben a una precisió del 83,9%.

Shulte im Walde (1998), (2000) ha investigat la classificació dels verbs en anglès i alemany mitjançant *clustering*, un procediment no supervisat que pretén "uncover an inherent natural structure of the data objects" (Schulte im Walde, 1998, pàgina 48). Per a l'anglès, emprà 80 classes de Levin per classificar 153 verbs a partir d'informació de subcategorització i de preferències de selecció, amb uns resultats de precisió del 61%. Per a l'alemany, Schulte im Walde i Brew (2002) van aplicar la mateixa tècnica per classificar 57 verbs en 14 classes.

2.3.5 Conclusions

En resum, en pocs anys s'ha avançat molt en el camp de l'adquisició lèxica des dels primers treballs de Brent (1991). El verb ha estat la categoria més tractada, sobretot per llengües com l'anglès i l'alemany. En concret, hi ha hagut molts treballs centrats en extreure de corpus informació sobre la subcategorització verbal. Els sistemes desenvolupats han arribat a nivells de precisió acceptables, tot i que encara hi ha marge per millorar.

La nostra aproximació s'allunya dels primers experiments, que pretenien recopilar de manera automàtica informació sobre els verbs i, en canvi, s'acosta més a la metodologia emprada en treballs més recents. Com Merlo i Stevenson (2001) definirem trets que puguin ser extrets automàticament. Emprarem dos mètodes diferents per classificar els verbs: un mètode

supervisat d'inducció de regles (com Merlo i Stevenson (2001)) i *clustering*, un mètode no supervisat, (com Schulte im Walde (1998)). A diferència dels treballs més recents d'adquisició verbal, no classificarem els verbs en classes semàntiques, sinó en classes sintàctiques.

Experiments amb mètode supervisat

3.1 Introducció

En aquest capítol presentem els experiments de classificació de verbs en transitius i intransitius amb un sistema supervisat. Inicialment, vàrem triar aquest mètode, ja que semblava l'aproximació més eficient, considerant que disposàvem d'una codificació manual de gairebé 8.000 verbs (v. secció 3.2.2). Els mètodes supervisats parteixen d'un conjunt d'objectes (lemes verbals, en el nostre cas) anotats i dedueixen regles que serviran per classificar-los en les classes establertes. Vàrem triar Ripper (Cohen, 2002), un *software* que indueix regles de classificació a partir d'un conjunt d'exemples prèviament classificats. Com a *input*, Ripper necessita un conjunt d'objectes, classificats, i descrits a partir de diversos trets. A la figura 3.1, podem veure la representació parcial de dos verbs: primer, hi ha el lema; després, la classe a la qual pertany; i finalment, el valor d'alguns trets (vegeu la secció 3.3 per a una explicació detallada dels trets).

contemplar	tr	9,3	52,2	3,4	4,3	15 ...
xisclar	intr	0	12,7	0	22	11 ...

Figura 3.1: Exemple de l'input per a Ripper

Ripper analitzarà els exemples i extreurà regles a partir de les dades per predir la classe d'altres objectes. Les regles establiran condicions que han de complir els objectes per ser classificats en una classe (per exemple, si un objecte té un valor superior a x pel tret y , classifica'l com a z).

3.2 Materials i mètodes

3.2.1 El corpus

Per tal de dur a terme els experiments d'adquisició de la subcategorització verbal, hem emprat una part del Corpus Textual Informatitzat de la Llengua Catalana (CTILC, d'ara endavant) (Rafel, 1994). El CTILC ha estat elaborat per l'Institut d'Estudis Catalans, conté textos escrits des de 1833 a 1988 i consta de 52.375.044 mots. Pretén ser un corpus representatiu de la

llengua i inclou textos de llengua literària (narrativa, teatre, poesia, assaig) i llengua no literària (tractats i manuals, articles en publicacions especialitzades, textos legals, premsa diària, etc.). El 56% dels mots corresponen a textos de llengua no literària i el 44% a llengua literària. A les taules 3.1 i 3.2 es veu la distribució de llengua literària i no literària segons el nombre de textos i de mots:

Gènere	% textos	% mots
assaig	12,3	13
narrativa	29,9	60
poesia	27,6	11
teatre	30,1	16

Taula 3.1: Distribució de textos literaris al CTILC, segons el gènere

Àrea	% textos	%mots
Filosofia	4	6
Religió i teologia	8,2	10,2
Ciències socials	15,6	19,1
Premsa	18,6	12,3
Ciències pures i naturals	6	7,6
Ciències aplicades	13,1	15,3
Belles arts, oci, esports	9,3	9,6
Llengua i literatura	9,3	7,6
Història i geografia	9,8	12
Correspondència	6,1	0,5

Taula 3.2: Distribució de textos no literaris al CTILC, segons l'àrea

La part del CTILC amb què s'ha fet aquest treball constava de 8 milions de mots quan vam iniciar. Posteriorment, hem pogut treballar amb un fragment major, que consta de més de 16 milions de mots. No disposem de dades sobre el percentatge de llengua literària i no literària que conté aquest fragment.

El CTILC és un corpus anotat: per a cada paraula, trobem el seu lema i una etiqueta amb informació morfològica (categoria principal i trets flexius). Va ser etiquetat semiautomàticament i corregit manualment i el seu etiquetari segueix l'estandàrd Eagles (1996). A 3.2 en tenim un exemple:

Per exemple, una etiqueta com *NCFSSO*. correspon a “nom comú femení singular” i *TDFSO* a “article determinat femení singular”.

mot	lema IEC	etiç. IEC
la	ell	PP3FSO..
veié	veure	VMIS3S.
amb	amb	SPS..
extrema	extrem	AQPFS0.
claredat	claredat	NCFS00.
i	i	CC00
molt	molt	RGP.0
precisa	precís	AQPFS0.
en	en	SPS..
la	el	TDFS0
seva	seu	DP3FS0.
imatge	imatge	NCFS00.
.	senselema	FE

Figura 3.2: Extracte del CTILC

3.2.2 Classificació de verbs

També disposem d'una taula amb informació de subcategorització anotada manualment de 7919 verbs. Aquesta taula va ser elaborada en el marc del Centre de Referència en Enginyeria Lingüística. Els verbs de la figura 3.3 en són exemples.

copiar	<S>	<O>	<NA>
acordar	<S>	<o>	<Ps> <C/INF> <C_que/INF_de> <NA>
marxar	<S>	<NA>	

Figura 3.3: Exemple de la codificació verbal

Les informacions que es codifiquen en les etiquetes són les següents:

- <S>: el verb pot portar objecte directe. <SS>: el verb no pot portar objecte directe.
- <O>: el verb porta objecte directe obligatori. <o>: l'objecte directe no és obligatori.
- <NA>: verb no atributiu. <A>: verb atributiu.
- <P>: verb que es pot pronominalitzar. <Ps>: el verb pot aparèixer pronominalitzat, cas en què no presenta objecte directe. <Po>: el verb pot aparèixer pronominalitzat i amb objecte directe alhora.
- <C>, <INF>, <C/INF>: el verb pot aparèixer complementat per una completiva, un infinitiu o ambdues construccions, respectivament.
- <C_que>, <C_que/si>, <INF_de> (o combinacions com <C_que/INF_de>): el verb pot portar completiva introduïda per *que*, per *que* o *si* o pot anar complementat per un infinitiu, respectivament.

Segons aquesta classificació manual, dels 7919 verbs que hi figuren, 6568 són transitius (un 82,9%) i 1351 són intransitius (17,1%). Hi ha 1291 verbs que apareixien més de cinquanta vegades en el fragment inicial del CTILC (8 milions de paraules): 1150 són transitius (89%) i 141 són intransitius (11%). Els verbs transitius inclouen els verbs opcionalment transitius (per exemple, *somiar*) o els que poden tenir alternances intransitives o de règim (per exemple, *fondre, parlar* o *referir*). Els verbs intransitius inclouen també els verbs de règim (com, per exemple, *contribuir*).

En el primer grup d'experiments realitzats es pretenia que el programa d'inducció de regles pogués aprendre a distingir entre verbs l'etiqueta dels quals contingués <S> o <SS>. És a dir, l'objectiu era aconseguir diferenciar verbs transitius i intransitius. En un primer moment, ens va semblar que també es podria fer servir la distinció entre <O> i <o> per detectar diversos tipus de transitius. Tanmateix, després d'uns experiments previs, aquesta distinció no s'ha fet servir ja que s'emprava de manera poc coherent. El principal problema és que, sota l'etiqueta <o>, s'hi agrupen verbs amb comportaments molt diferents: verbs *object-drop*, com *telefonar*; verbs causatius, com *tancar*; verbs intransitius amb algun ús transitiu, com *plorar* i verbs que apareixen majoritàriament amb objecte directe explícit, com *explicar*. Sota l'etiqueta <O>, també podem trobar verbs que sovint prescindeixen de l'objecte directe, com *cavil·lar* o *mastegar*. La informació sobre complementació amb completives i infinitius no s'ha fet servir, tot i que podria ser útil, en etapes posteriors, per ampliar els experiments que presentem en aquest treball.

3.3 Definició de trets

En aquest apartat, explicarem els trets que proposem per distingir entre verbs transitius i verbs intransitius i que seran la base dels experiments per a aquest capítol i el capítol següent.

Com hem explicat en el capítol 2, en el treball de Merlo i Stevenson (2001) es considerava que un verb tenia el tret de transitivitat si anava seguit d'un número, pronom, determinant, adjectiu o nom. En canvi, es comptava com a intransitiu si anava seguit de signe de puntuació, conjunció, preposició, partícula o data. El fet que les llengües romàniques, a diferència de l'anglès, tinguin un ordre sintàctic molt més flexible fa que sigui més difícil extreure dades fiables. Mentre que, en anglès, un nom darrere el verb té moltes possibilitats de ser l'objecte, en català tant pot ser l'objecte com el subjecte. A més, en català, l'objecte tampoc va sempre just darrere del verb, sinó que pot anar darrere, però molt separat (per un circumstancial, per exemple, cosa molt menys freqüent en anglès) o davant del verb. Evidentment, aquesta

flexibilitat sintàctica dificulta l'extracció de dades.

Hem establert deu trets que ens serviran per fer la distinció entre transitius i intransitius. Els deu trets establerts defineixen correlats lingüístics superficials que es poden detectar automàticament en el corpus CTILC amb un grau alt de fiabilitat. Per tant, s'aprofitarà la mena d'informació amb què el CTILC és anotat: principalment, informació sobre lema i sobre categoria morfològica (tant categoria major -nom, verb-, com trets flexius -singular, femení). Els deu trets són els següents (per a cada tret, incloem un exemple en què el verb té el tret que estem explicant¹):

DirClitic: Verb seguit o precedit per un pronom feble de complement directe: *el, la, les, ho*. No cercarem aquells pronoms que també puguin fer funció de complement indirecte (com *els*). Caldria esperar que els verbs transitius presentin valors molt majors per a aquest tret que els verbs intransitius.

(29) Si és vedella, els pagesos la **guarden** per a la cria i la llet

DetONom: Verb seguit de determinant o nom. Els verbs transitius haurien de presentar valors alts per aquest tret. Tanmateix, sembla probable que alguns verbs intransitius presentin també un valor molt alt en aquest tret, ja que és molt freqüent que certs verbs intransitius (els inacusatius com *arribar, aparèixer*) apareguin amb el subjecte postverbal (Rosselló, 2002), (Vallduví, 2002). Tant el verb *assenyalar* com el verb *aparèixer* tindrien aquest tret en els exemples següents.

(30) a. No hem trobat cap rastre ni cap ressò que **assenyali** la viabilitat d'alguna pista

b. Han **aparegut** unes màquines per a tallar el pernil molt prim, més aviat translúcid, subtil i delicat

Passiva: Un verb tindrà aquest tret si es troba en un dels contextos següents:

- Verb en participi precedit del verb *ésser*.

(31) Una part dels seus béns seran **devorats** per les flames

¹Si no s'indica el contrari, tots els exemples pertanyen al CTILC.

- Verb en participi seguit de la preposició *per*. Tant aquest context com l'anterior detecten la forma passiva clàssica.

(32) Sortia únicament quan se sentia **empaitat** per la fam i la set

- Verb precedit del pronom *se* seguit de nom². Aquest tret detecta passives reflexives

(33) Així se **seleccionen** plantes en què el vent deixa de ser el principal agent de dispersió

Els verbs intransitius haurien de mostrar valors molt més baixos que els transitius.

Punt: Verb seguit d'un dels següents signes de puntuació: punt, punt i coma, dos punts, interrogació o exclamació. Els verbs intransitius haurien de mostrar valors més alts per a aquest tret.

(34) Tecleta **riu**. Jo el **renyo**.

Prep: Verb seguit de preposició, excepte la preposició *per*. Excloent la preposició *per*, evitem comptar aquelles ocurrències en què la preposició introdueix el complement agent. Els verbs intransitius, que en la nostra classificació inclouen els de règim, haurien de mostrar valors més alts que els transitius per a aquest tret

(35) Era com si **flotés** en un mar profund i sense ribes perceptibles

Se: Verb seguit o precedit del pronom *se*. Com hem explicat a la secció 2.2.3, aquesta partícula apareix molt freqüentment incrementant al verb. Sembla probable que la seva freqüència sigui major amb verbs transitius ja que només els verbs transitius poden ser usats de forma reflexiva, recíproca o poden patir operacions de passivització (situacions en les quals apareix aquesta partícula). Tanmateix, també apareix amb verbs intransitius (com, per exemple, *morir-se*).

(36) a. A tot això se **sacrifiquen** les “distincions” d'altres temps

b. Com més es **creix**, més sòlida és la base

²Agraeixo al Toni Badia la idea per aquest context.

Els últims quatre trets que presentem tenen per objectiu distingir entre verbs transitius i verbs inacusatius (o verbs intransitius amb subjecte postverbal). Si detectem un nom o determinant darrere d'un verb, no podem estar segurs de si es tractarà d'un objecte o d'un subjecte. Tanmateix, si restringim més el context d'aparició (controlant la concordança, la presència simultània d'objecte i subjecte o la forma del verb), sí que ens podrem assegurar amb més fiabilitat que només els verbs transitius tinguin aquest tret, ja que haurem exclòs els verbs intransitius amb subjecte postverbal.

DetAmbNom: Verb seguit de determinant o nom i precedit d'adjectiu, pronom fort, determinant o nom. Aquest tret preten detectar alhora un potencial objecte darrere el verb i un subjecte potencial davant del verb (un nom, un adjectiu, un pronom fort, etc.). Per tant, si un verb té aquest tret, probablement, no es tractarà d'un intransitiu amb subjecte postverbal, sinó d'un transitiu.

(37) El seu esguard **adquirí** un aire ombrívol

NoCordança: Verb seguit d'un determinant o nom amb el qual no concorda, amb la qual cosa restringim molt la possibilitat que es pugui tractar d'un subjecte.

(38) **Menjà** aliments infectes, **prengué** beuratges amargs

VerbNoPersonal: Verb en forma no personal (infinitiu, participi o gerundi) seguit de determinanat o nom.

- (39) a. El mar apareixia rutilant i d'un blau exaltat i compacte, **desvetllant** una alenada d'alegria i d'esperança
- b. Un cop **arribat** el producte a nivell de l'efector, cal que interactuï amb unes determinades molècules

Segons Bel (2002), les formes no finites del verb es caracteritzen per no poder funcionar com a verbs principals i estan incapacitades per portar subjectes explícits, en el cas general. Tanmateix, en algunes construccions concretes, sí que podem trobar verbs en forma no personal amb subjecte postverbal (com l'exemple 39b). De totes maneres, sembla que si trobem un verb amb forma no personal seguit de nom o determinant, la probabilitat que es tracti de l'objecte és molt més gran que si el verb estigués flexionat. Per poder confirmar aquesta hipòtesi hem extret

dades d'alguns verbs transitius i intransitius triats a l'atzar. Concretament, hem calculat dos percentatges: el percentatge d'ocurrències en què cada verb està seguit de determinant o nom (DN) i el percentatge d'ocurrències en què el verb està en forma no personal (VNP) i seguit de DN. S'han descartat les ocurrències en què el verb anava precedit d'un altre verb per tal de no tenir en compte les formes compostes del verb (*va néixer, ha iniciat*). Podem veure els resultat a la taula 3.3.

Verb	Num. oc.	% DN	% VNP DTN
Ballar	736	14,2	5,7
Pintar	639	23,6	7,8
Transmetre	525	28,1	12,5
Iniciar	1753	48,3	12,8
Protegir	816	26,5	13,4
Nedar	215	6	1,3
Costar	1059	14,2	0,28
Somriure	1473	2,7	0,2
Néixer	2175	14,2	1,6
Ocórrer	426	7,5	0,4

Taula 3.3: Percentatge d'ocurrències seguides de DN i d'ocurrències VNP + DN. Els cinc primers verbs són transitius i els cinc darrers són intransitius

Podem observar que, en general, el percentatge d'ocurrències seguides de DN és major per als verbs transitius que per als intransitius. Tanmateix, per alguns intransitius (com *costar* o *néixer*), el percentatge és també força alt, comparable al d'alguns verbs transitius (com *ballar*³). En canvi, la mitjana de VNP + DN és molt baixa per a tots els intransitius (en cap cas arriba al 2%). Per tant, si trobem la seqüència VNP + DN es tractarà molt probablement d'un verb transitiu. En canvi, si trobem la seqüència Verb + DN tant pot ser un verb transitiu com un inacusatiu.

V12: Verb en primera o segona persona seguit de determinant o nom, amb la qual cosa el sintagma que segueix el verb no serà, molt probablement, el subjecte⁴. En aquests casos, probablement es tractarà de l'objecte (frase 40a), tot i que també ens podríem trobar amb, per exemple, complements circumstancials de temps (frase 40b). Per evitar detectar casos semblants als de 40b, hem elaborat una llista amb els noms temporals més freqüents (com, per exemple, *dia* o *Nadal*), de manera que, si apareix un d'aquests noms seguint el verb, no es tindria en compte aquesta ocurrència del verb.

³Recordem que, segons la nostra classificació, *ballar* és transitiu, ja que admet complement directe, si bé s'empra molt sovint sense.

⁴En són excepcions expressions com *venim els catalans d'un país (...)*. Tanmateix, són construccions molt poc freqüents estadísticament, amb la qual cosa el tret segueix sent vàlid

(40) a. Aquí **citem** un fragment de l'himne segons l'edició de Migne

b. I **sortíem** cada any amb la guardiola per la Creu Roja i el càncer

Per extreure els valors dels trets per a un lema, cal comptar quantes de les seves ocurrences presenten cada tret i fer el percentatge entre el nombre total d'ocurrences. Per exemple, el verb *xisclar* apareix 119 vegades en el corpus, de les quals 14 va seguit de determinant o nom. Per tant, el valor del tret DetONom pel verb *xisclar* és 11,7%.

A la figura 3.4, podem veure la representació completa dels verbs *contemplar* i *xisclar*. El verb *contemplar*, per exemple, va seguit o precedit de clíctic de complement directe (tret DirCLitic) en un 9% dels casos, mentre que *xisclar* en un 0%. *Xisclar* va seguit de signe de puntuació (tret Punt) en un 22% dels casos i *contemplar* en un 4%.

Lema	Etq.	DirCLitic	DetONom	Passiva	Punt	Prep
contemplar	tr	9,3	52,2	3,4	4,3	15
xisclar	intr	0	11,7	0	22	11
		Se	DetAmbNom	NoConcordança	VerbNoPersonal	V12
contemplar	tr	5,9	15,1	17,3	25,4	13,7
xisclar	intr	0,8	0	0	6,6	0

Figura 3.4: Valors dels trets per als verbs *contemplar* i *xisclar*

A la taula 3.4, podem veure les mitjanes dels valors dels trets per a cada categoria, les quals s'ajusten aproximadament a les nostres previsions:

- Els verbs transitius tenen una mitjana clarament més elevada que els intransitius per als trets: DirCLitic, DetONom, Passiva, Se, NoConcordança, VerbNoPersonal i V12. Els transitius tenen una mitjana lleugerament més elevada per al tret DetAmbNom.
- Els verbs intransitius tenen una mitjana més elevada per als trets Prep i Punt.

Tanmateix, cal observar que els lindars són força baixos per a tots els trets (només DetONom i Prep mostren percentatges relativament alts). Aquest fet, com veurem, dificultarà la deducció de regles.

3.4 Experiments i resultats

L'*input* per als experiments és un fitxer amb els 1291 verbs anotats tal i com vèiem a la figura 3.4. A partir d'aquestes dades, Ripper (Cohen, 2002), provarà de fer generalitzacions sobre el comportament de cada classe.

Tret	Tots	Transitius	Intransitius
DirClitic	4,3	4,5	0,1
DetONom	35,7	38,0	16,6
Passiva	5,1	5,7	0,3
Punt	7,8	7,6	9,5
Prep	22,9	20,9	39,8
Se	14,3	15,7	2,5
DetAmbNom	15,3	15,5	13,9
NoConcordança	14,9	16	5,5
VerbNoPersonal	17,7	18,7	9
V12	5,6	6	1,8

Table 3.4: Valors dels trets per als transitius i per als intransitius

Es van dur a terme diferents experiments seguint dues metodologies diferents: la basada en entrenament (*test*) i avaluació (*train*) i la basada en el mètode de *cross-validation*.

Per a la primera metodologia, vàrem dividir els verbs en dos grups a l'atzar: un grup d'entrenament (GE, d'ara endavant) i un grup d'avaluació (GA, d'ara endavant). El GE conté el 90% dels objectes (1162 verbs) i el GA el 10% (129 verbs). L'algoritme té en compte només les dades del GE per dur a terme les seves prediccions i deduir regles. Després aplica aquestes regles sobre els verbs del GA, verbs que no havia emprat per deduir les regles. El fet d'emprar dos grups diferents de verbs ens permetrà descobrir si l'algoritme també és útil per classificar verbs que no havia emprat per deduir les regles. Presentem els resultats de l'experiment a la figura 3.5

```

Final hypothesis is:
intr :- DirClitic ≤ 0,383632, Se ≤ 0,866739 (73/17).
intr :- DirClitic ≤ 0,215661, Se ≤ 6,72269 , NoConcordança ≤ 3,4965 (21/2).
intr :- Passiva ≤ 0,178731, DetONom ≤ 12,3249, Se ≤ 2,60304 (7/4).
intr :- DirClitic ≤ 0,10929, VerbNoPersonal ≤ 4,16667, Se ≤ 30,9008 (9/5).
default tr (1013/11).
===== summary =====
Test error rate: 3.36% +/- 0.53% (1162 datapoints)
Train error rate: 4.65% +/- 1.72% (129 datapoints)
Hypothesis size: 4 rules, 15 conditions
Learning time: 0.64 sec

```

Figura 3.5: Resultats de Ripper amb entrenament i avaluació

La taxa d'error en els objectes del GE (*test error rate*) és del 3,36% i en els objectes del GA (*train error rate*) del 4,65%. L'algoritme estableix que, per defecte, un verb serà transitiu (*default tr*) i dedueix quatre regles per detectar verbs intransitius. Segons aquest experiment, el verb serà intransitiu si:

- El valor de DirClitic és menor de 0,38 i el valor de Se menor de 0,86. Aquesta és la regla més emprada: detecta correctament 73 verbs i erròniament 17 verbs.
- El valor de DirClitic és menor de 0,21, el valor de Se és menor de 6,72 i el valor de NoConcordança és menor de 3,49. Aquesta regla detecta correctament 21 verbs intransitius i erròniament 2 verbs.
- El valor de Passiva ha de ser menor de 0,17, el de DetONom menor de 12,32 i Se menor de 2,6. Aquesta regla detecta 7 verbs intransitius i erròniament 4.
- El valor de DirClitic és menor de 0,1, el valor de VerbNoPersonal menor de 4,16 i el valor de Se menor de 30,9. Aquesta regla detecta correctament 9 verbs i erròniament 5 verbs.
- L'opció per defecte detecta correctament 1013 verbs i erròniament 11.

Les regles empren 6 dels 10 trets. Tots es fan servir de la manera esperada. El tret Se es fa servir en totes les regles, cosa que ens pot donar una idea de la importància que té aquest tret com a element distintiu entre transitius i intransitius. La segona i, especialment, la primera regla són les més utilitzades; mentre que la tercera i la quarta es fan servir en pocs casos. No es fan servir els trets Punt, Prep, DetAmbNom i V12. El tret Prep no es fa servir en aquest experiment concret, però, com veurem amb els resultats amb *cross-validation* és un tret útil per distingir transitius i intransitius. Pel que fa a Punt, DetAmbNom i V12, són els tres trets que mostren valors mitjans més semblants per als verbs transitius i intransitius (v. taula 3.4) i, per tant, és més esperable que siguin menys importants. Tanmateix, aquests trets, com veurem més endavant, tot i que no apareguin en les regles, serveixen per reforçar la distinció entre transitius i intransitius.

En la metodologia *10-fold cross-validation*, es divideixen els objectes en deu parts a l'atzar i s'aplica l'algoritme d'aprenentatge deu vegades sobre diferents particions de verbs en GA i GE (el GA representa sempre el 10% i el GE el 90%). Al final, s'obté la mitjana de la taxa d'error que s'ha obtingut en cada una de les deu execucions de l'algoritme sobre les dades. També s'obté el conjunt de regles que extreu l'algoritme per a cada una de les deu vegades que s'executa (adjuntem el conjunt complet de regles a l'annex A).

Fent servir aquesta metodologia s'obté una taxa d'error mitjana del 4,56%. A la taula 3.6, presentem els resultats que obté per a una de les particions, en què la taxa d'error és del 3,4%. En aquesta partició, l'algoritme dedueix cinc regles fent servir vuit trets diferents. Les regles s'ajusten força amb les nostres previsions.

Final hypothesis is:
 intr :- DirClitic \leq 0,311526, Se \leq 4,91453, NoConcordança \leq 3,38983 (47/5).
 intr :- DirClitic \leq 0,518135, Se \leq 0,518135 (29/6).
 intr :- DirClitic \leq 0,110988, Se \leq 5,78947, DetONom \leq 25,6927 (19/8).
 intr :- Passiva \leq 0,675676, Prep \geq 33,3333, V12 \leq 1,92308, Punt \geq 5,95238 (7/3).
 intr :- DetONom \leq 1,3742 (2/1).
 default tr (914/5).
 Error rate on holdout data is 3.44828%

Figura 3.6: Resultats de Ripper amb cross-validation

Tot els trets s'utilitzen com a mínim una vegada en alguna regla d'alguna de les deu particions. Alguns trets només s'utilitzen una vegada (V12 i DetAmbNom) i d'altres apareixen en totes les particions (Se i DirClitic).

Per conèixer la contribució individual de cada tret en la deducció de regles, hem fet els mateixos experiments eliminant cada vegada un dels deu trets emplatats. La taula 3.5 mostra la taxa d'error que s'obté tant amb la metodologia d'entrenament i avaluació com emprant *cross-validation*. La primera fila mostra els resultats finals, sense eliminar cap tret: 4,65% amb entrenament i avaluació i 4,56% amb *cross-validation*.

Tret eliminat	Entrenament i avaluació	Cross-validation
	4,65	4,56
DirCLitic	14,73	5,94
DetONom	6,98	5,34
Passiva	5,43	5,94
Punt	6,20	5,25
Prep	4,75	6,03
Se	7,75	6,80
DetAmbNom	6,20	5,42
NoConcordança	5,43	5,94
VerbNoPersonal	7,75	5,83
v12	8,53	6,68

Taula 3.5: Taxes d'error eliminant trets

La taxa d'error augmenta en tots els casos, sigui quin sigui el tret eliminat, tant fent servir la metodologia d'entrenament i avaluació, com emprant *cross-validation*. Amb aquesta última metodologia, l'augment de la taxa d'error és semblant per a tots els trets: entre un 0,69%, en el cas de Punt (passa d'un 4,56% a un 5,25%), i un 2,24%, en el cas de Se (passa d'un 4,56% a un 6,80%). Amb entrenament i avaluació, la variació és major: trobem augments d'entre 0,10%, en el cas de Prep (passa d'un 4,65 a un 4,75%) i un 10,08%, en el cas de DirClitic (passa d'un 4,65% a un 14,73%). Aquest augment tan fort que es produeix si eliminem DirClitic segurament es deu a les característiques concretes d'aquesta partició de verbs en GA i GE.

El fet que la taxa d'error augmenti sigui quin sigui el tret eliminat mostra que tots els trets són pertinents i vàlids per establir el grau de transitivitat (o intransitivitat) d'un verb. Fins i tot en el cas que un tret no aparegui en les regles (com era el cas de Punt, Prep, DetAmbNom i V12 en l'experiment amb entrenament i avaluació (figura 3.5)), el tret contribueix en certa manera a reforçar la descripció del verb i a facilitar-ne la classificació.

Així, els resultats amb les dues metodologies són molt semblants (4,65% vs 4,56%). És especialment positiu que els resultats amb *cross-validation* siguin lleugerament millors, ja que és una metodologia més fiable. L'algoritme apren deu vegades amb conjunts de verbs diferents, amb la qual cosa els resultats no depenen del conjunt concret de verbs del GE i del GA i, per tant, no poden ser producte de l'atzar, cosa que no passa amb la metodologia d'entrenament i avaluació. La metodologia d'entrenament i avaluació, però, resulta molt útil ja que, com a *output*, produeix una llista dels verbs del GA classificats segons les regles que ha deduit. Això ens permet conèixer els verbs pels quals les regles fan prediccions errònies, casos que analitzarem a la següent secció.

3.5 Anàlisi de resultats

3.5.1 Anàlisi d'errors

Dels 129 verbs del GA emprats amb la metodologia d'entrenament i avaluació, n'hi ha 6 (4,65%) en què la classificació manual i la predicció que fa l'algoritme no coincideixen. Els podem veure en la figura 3.7. En tots els casos, es tracta de verbs intransitius segons la nostra classificació manual que han estat classificats com a transitius. No hi ha cap cas de verb transitiu que hagi estat classificat com a intransitiu.

esdevenir
esmorzar
influir
morir
soler
transitar

Figura 3.7: Verbs mal classificats en el GA

A continuació, analitzarem breument perquè aquests verbs no han estat adequadament classificats.

- *Esmorzar* (250 ocurrències): Les regles no s'apliquen perquè els valors dels següents trets superen el llindar establert:

- Se. Només hi ha 2 casos en tot el corpus, però tanmateix ja se supera el llindar. Per exemple: *Es pot esmorzar a algun poble.*
- NoConcordança. Presenten aquest tret occurrències com, per exemple, *vam esmorzar pa amb formatge.* L'ús transitiu d'aquest verb, tot i no ser normatiu, apareix algunes vegades en el corpus.
- DetONom. El valor supera lleugerament el llindar imposat a la regla. Per exemple, tenen aquest tret els verbs de les següents frases: *Vaig esmorzar cafè, has esmorzar un entrecot.* Es tornen a detectar la mena d'estructures que hem vist en el tret NoConcordança.
- VerbNoPersonal. Només trobem dos casos en el corpus: *Ens convidaren a esmorzar pa amb tomaca; esmorzar pa i raïm.*

Les regles no s'apliquen perquè se supera per molt poc el valor establert a les regles. Aquest verb presenta alguns usos transitius, però no són majoritaris (aproximadament 8 occurrències de 250 en el corpus). Tanmateix, com que els llindars a les regles són molt baixos, els usos transitius prenen més rellevància de la que haurien de tenir.

- *Influir* (715 oc.): Aquest verb té alguns usos transitius, per la qual cosa alguns trets tenen uns valors massa alts que impedeixen que les regles s'apliquin, com ja passava amb *esmorzar*.
 - DirClitic. *Ja sé que a vostè aquestes coses no l'influeixen massa*
 - NoConcordança. *I abans, encara, Gassendi que tan influí els científics valencians, podria considerar-se (...)*
 - Passiva. *Els nostres estómacs han estat influïts en bona part pels aranzels de les Duanes.*

El DIEC (1995) no recull aquests usos transitius i només dona compte d'*influir* com a verb intransitiu.

- *Morir* (3471 oc.): Presenta valors molt alts en els següents trets:
 - Se. *Gent que es mor d'amor*
 - Passiva. *És morta no fa gaire; quan ells siguin morts.*
 - DetONom. *Havien mort els hiverns; es moren les preguntes.*
 - VerbNoPersonal. *Mort el dictador; morint l'embat lleuger; una vegada morta la flor*

- *Soler* (2001 oc.): Les regles no s'apliquen perquè aquest verb presenta valors alts en els dos dels trets que més importància tenen segons les regles:
 - Se. *Bastons que se solen posar, se solen produir confidències*
 - DirClitic. *El solien penjar, la sol fer, la solc respectar, hom la sol trobar*. En aquest cas, podem veure com el tret no era prou restrictiu i s'han detectat casos en què el clíctic complementa no el verb que volíem analitzar, sinó l'infinitiu que el segueix.
- *Transitar* (82 oc.): Aquest verb té alguns usos transitius, que, altre cop, no són gaire significatius, però fan que els valors dels trets sobrepassin els llindars establerts en les regles:
 - DirClitic. *Que l'havien transitada*. Només hi ha un sol cas en tot el corpus.
 - Se. *Per on es transita más; quan es transite de nit*.
 - NoConcordança. Només trobem un sol cas en tot el corpus. *Són pocs els qui transiten el carrer*
 - Passiva. *Els camins (...) eren transitats periòdicament per caravanes; no haver estat transitat per ningú*.
 - VerbNoPersonal. *Transitar aquell carrer*. De nou, només hi ha un sol cas en tot el corpus que tingui aquest tret.

Com en el cas d'*esmorzar*, el fet apareguin alguns usos transitius d'aquest verb, tot i que molt minoritaris en el corpus, fa que els trets que indiquen transitivitat (DirClitic, Passiva, etc.) tinguin valors massa alts i, per tant, les regles per detectar intransitius no s'apliquen.

- *Esdevenir* (4449 oc.): Aquest verb, com *parèixer* i *semblar*, està classificat com a <SS><A>, és a dir verb atributiu que no pot portar objecte directe. Per als nostres experiments no hem previst la classe de verbs atributius, sinó que només hem distingit transitius i intransitius. La primera part de l'etiqueta dels tres verbs atributius és <SS>, és a dir, l'etiqueta que correspon als verbs intransitius. Tanmateix, aquest no és un verb intransitiu, com tampoc no és transitiu. Per tant, no es trobem davant d'un contraexemple, sinó d'una classe que no hem tingut en compte, ja que és molt poc productiva, amb la qual cosa no té gaire sentit intentar adquirir-la. L'algoritme classifica *esdevenir* com a transitius, ja que supera els llindars establerts per alts trets DetONom i Se. Els verbs atributius són superficialment molt semblants als transitius, ja que subcategoritzen un complement (v. exemple 41) i, per tant, no és estrany que així l'hagi considerat l'algoritme.

(41) El dia esdevé una pura il·lusió de l'esperit, insignificant

3.5.2 Avaluació

Els resultats finals presenten una taxa d'error del 4,56%. Tot i ser uns bons resultats, cal tenir en compte que si classifiquéssim atorgant a tot objecte la classe per defecte (transitiu), la taxa d'error seria tan sols un 6,44% més alta, de l'11% (percentatge de verbs intransitius en la classificació manual).

L'algoritme d'aprenentatge és capaç d'extreure algunes generalitzacions a partir dels valors dels trets i de fer algunes regles útils. No obstant això, totes les regles que extreu presenten uns llindars bastant baixos. És a dir, les condicions per ser un verb intransitiu són molt exigents. Si un verb intransitiu presenta un valor lleugerament superior a zero en alguns dels trets que defineixen els transitius (DirClitic, Passiva, NoConcordança, etc.) les regles no s'aplicaran i s'etiquetarà com a transitiu, ja que és la classe per defecte. Això és el que passa amb verbs com *esmorzar* o *transitar*. Com que *esmorzar* apareix en un ús transitiu en certes ocurrències del corpus (que, ni molt menys, són majoritàries), això provoca que les regles no es puguin aplicar.

L'algoritme es veu forçat a posar aquests llindars tan baixos a causa de la classificació que s'ha emprat. En aquesta classificació, si un verb té un ús possible com a verb transitiu, per molt marginal que sigui aquest ús, és classificat com a verb transitiu. Així, verbs com *avenir*, *al.ludir*, *dormir*, *trucar* o *acostumar* apareixen com a transitius. Tanmateix, en el cas de *dormir*, per exemple, només 7 ocurrències de les 1785 que té en el corpus presenten un marc de subcategorització transitiu (*dormir la sesta* (3 vegades), *dormir la turca*, *dormir el son de l'oblit*, *dormir la bufa* i *dormir la mona*). Tot i que les altres 1778 ocurrències són intransitives, aquest verb està classificat com a transitiu.

Aquests verbs, tot i estar classificats com a transitius, tenen un ús majoritàriament intransitiu i, per tant, els valors dels trets seran més semblants als dels verbs intransitius que als dels verbs transitius. Això fa que les diferències entre transitius i intransitius es "difuminin", siguin molt menys clares i, per tant, l'algoritme ha de posar condicions molt estrictes per tal que un verb pugui ser considerat intransitiu.

És a dir, aquesta classificació té un concepte de transitivitat molt ampli i un concepte d'intransitivitat molt restringit. Recordem que només l'11% dels verbs classificats estaven considerats com a intransitius, cosa que fa més difícil la deducció de regles.

Per tant, la classificació emprada no s'adequa a les nostres necessitats, ja que estem fent

servir mètodes estadístics per adquirir informació lèxica i, en canvi, la classificació no té en compte un factor clau: la freqüència d'ús del verb en un marc argumental o en un altre. Evidentment, la freqüència d'un verb en un marc argumental és clau per a que els algoritmes d'aprenentatge automàtic puguin adquirir correctament la classe del verb. A més, és un factor important no tan sols des del punt de vista tècnic, sinó també des del punt de vista lingüístic. Si un verb apareix en un marc intransitiu el 99,6% de les ocurrències d'un corpus i en un marc transitiu el 0,4% (com passava amb *dormir*), és evident que aquesta informació és rellevant alhora d'analitzar-lo lingüísticament.

Una possible via per resoldre la inadequació de la classificació manual hauria estat revisar-la adequant-la als nostres criteris. Tanmateix, això resoldria el problema només parcialment, ja que el sistema continuaria depenent d'una classificació i, per disposar d'una classificació, sovint s'han de prendre decisions més o menys arbitràries, especialment quan cal tractar verbs poc prototípics, conflictius o que es troben a la frontera entre les classes. Per això, hem considerat que el més adequat era canviar de mètode i emprar-ne un de no supervisat, que no depengui d'una classificació prèvia i d'uns criteris preestablerts i que, per tant, presenti més garanties empíriques.

Una de les aportacions d'aquest capítol és la definició de trets que presenten correlats lingüístics superficials i que estableixen el grau de transitivitat (o intransitivitat) d'un verb. Així, si un verb té valors alts pels vuit trets que indiquen transitivitat (DirClitic, DetONom, Passiva, Se, DetAmbNom, NoConcordança, VerbNoPersonal, V12) i valors baixos pels dos trets que indiquen intransitivitat (Punt i Prep), es tractarà d'un verb molt prototípicament transitiu. Si, en canvi, un verb mostra el patró contrari, es tractarà d'un verb molt clarament intransitiu. Entre aquests dos casos tan clars, hi ha molts casos intermitjos, que s'allunyen gradualment dels prototipus i que mostren comportaments més difícils de classificar clarament en una categoria tancada.

3.6 Conclusions

En aquest apartat, hem exposat els experiments realitzats amb un mètode supervisat destinats a distingir entre verbs transitius i verbs intransitius.

Els resultats són força acceptables i el sistema ha estat capaç d'adquirir regles per classificar verbs. El millor resultat que hem obtingut mostrava un taxa d'encert del 95,44% (taxa d'error del 4,56%). Tanmateix, ens hem adonat que la classificació no s'adequa a les nostres necessitats, per la qual cosa els resultats i les prediccions que fa l'algoritme estan esbiaixats. En

conseqüència, hem decidit que és millor deixar de fer servir un mètode supervisat i que és més convenient emprar-ne un de no supervisat, que no depengui de cap classificació. Presentem els experiments fets amb un mètode no supervisat, el *clustering*, en el següent capítol.

Experiments amb *clustering*

4.1 Introducció

En aquest capítol, presentem els experiments realitzats amb un mètode no supervisat, el *clustering*. Emprant aquesta tècnica, l'algoritme no s'ha d'amotllar a una classificació prèvia, sinó que analitza els valors dels trets de cada verb i , a partir d'aquestes dades, els divideix en grups, en *clusters*. El programa triat per dur a terme aquests experiments ha estat Cluto (Karypis, 2002). El que fa aquest programa és

divide data into meaningful or useful groups, called *clusters*, such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized. (Karypis (2002), pàgina 1).

Aquesta tècnica s'ha fet servir per coses tan diferents com classificar els documents de la Web, caracteritzar diferents tipus de consumidors segons el tipus de compres que fan, agrupar els gens i les proteïnes que tenen una funció semblant o agrupar llocs que tenen una probabilitat similar de patir terratrèmols. En el nostre cas, volem descobrir tipus de verbs que tenen un comportament sintàctic semblant.

L'*input* per a aquests experiments és el mateix que l'*input* dels experiments que s'han fet amb el programa d'inducció de regles (vegeu el capítol 3); és a dir, els verbs descrits a partir dels valors de 10 trets que hem proposat per distingir diferents comportaments i , per tant, diverses classes. La diferència és que, en aquest cas, els verbs no estan classificats i , per tant, aquesta no és una informació que es tingui en compte per fer els grups.

4.2 Materials i mètodes

4.2.1 Classificació dels verbs

Com hem explicat, el *clustering* ofereix com a resultat els objectes (els verbs) agrupats en diversos *clusters*. Avaluar els resultats manualment analitzant cada *cluster* resulta lent i difícil.

Per això, per tal de poder avaluar els resultats de l'experiment, hem seleccionat 200 verbs a l'atzar (dels 1291 que apareixien més de 50 vegades en el corpus) i els hem classificat manualment en tres classes:

- Verbs transitius: Verbs que admeten la passiva (tot i que hi ha excepcions com *tenir*) i apareixen amb objecte directe, tot i que també es puguin emprar sense (*aguantar*).
- Verbs intransitiu: Verbs que no admeten complement directe. També hi incloem verbs que exigeixen un complement de règim verbal i aquells que tenen algun ús transitiu molt poc significatiu estadísticament: per exemple, *trucar* (*trucar la moto*) o *dormir* (*dormir la mona*).
- Verbs d'alternança amb *se* (VASE, d'ara endavant): Verbs que presenten la següent alternança:
 - Quan no apareixen en forma pronominal, requereixen complement directe: *aprofitar una oportunitat*.
 - Quan apareixen en forma pronominal, solen aparèixer amb un complement preposicional: *aprofitar-se d'una persona*.

Dins la classe VASE, trobem dos patrons d'alternança diferents (n'hem parlat en les seccions 2.2.1.3 i 2.2.2):

- Verbs que presenten una veritable alternança entre un complement directe i un complement de règim.
 - (42) a. Sobretot si la gresca **beneficiava** la vila d'aquella manera tan inesperada com espectacular
 - b. També el grup es **beneficia** d'un augment de capacitat d'adaptació
- Verbs que, en la seva forma no reflexiva, apareixen amb un complement directe i un complement preposicional. Amb la forma reflexiva, el directe desapareix i el preposicional es manté.
 - (43) a. El que **diferencia** l'home de qualsevol altre animal és que produeix els seus mitjans de vida material
 - b. Sense aquestes limitacions, no es podria **diferenciar** dels polítics

Hem decidit establir aquesta tercera classe de verbs (a part dels transitius i intransitiu) per diverses raons:

- Els verbs d'aquesta classe es troben a mig camí de les altres dues classes definides: transitius i intransitius. En el corpus, apareixen tant en construccions transitives com en intransitives.
- És possible detectar-los fent servir informació purament morfològica, ja que presenten uns correlats lingüístics superficials clars: la partícula *se* i preposició en l'alternança intransitiva; el complement directe en l'alternança transitiva.
- Saber que un verb pertany a aquesta classe ens dóna informació útil per, per exemple, fer anàlisi sintàctica automàtica. Si el nostre analitzador troba un verb que pertany a la classe VASE, sabrà que, si apareix acompanyat de la partícula *se*, no anirà seguit d'un complement directe, sinó probablement d'un preposicional. Si, en canvi, no apareix aquesta partícula, probablement anirà seguit d'un complement directe.

Els 200 verbs classificats apareixen a la figura 4.1.

1. **Transitius** (129, 64,5%): abaixar, admetre, administrar, adorar, afirmar, afrontar, aguantar, alterar, ampliar, animar, apagar, aprofundir, arreglar, assenyalar, atacar, atendre, atorgar, atreure, autoritzar, avorrir, buidar, bullir, capgirar, causar, celebrar, cessar, cloure, comercialitzar, commoure, concebre, conèixer, conferir, configurar, connectar, consultar, contrastar, coure, curar, demanar, demostrar, denominar, depassar, destruir, detallar, dur, edificar, eixamplar, elaborar, elegir, encaçar, enfocar, enviar, errar, esborrar, esclafar, escoltar, espisar, espolsar, estimular, estirar, evocar, examinar, exceptuar, felicitar, finalitzar, finançar, foradar, forjar, fornir, heretar, il.luminar, imitar, inaugurar, iniciar, insinuar, intensificar, intercanviar, jurar, mesurar, minvar, netejar, obeir, observar, odiar, operar, ordenar, pensar, pescar, pintar, posseir, prendre, presidir, pressentir, programar, protegir, provocar, puntualitzar, realitzar, recórrer, rectificar, reular, refer, refusar, regar, regirar, regular, reivindicar, renovar, reparar, reposar, representar, ressuscitar, retrobar, revelar, sacsejar, satisfer, seleccionar, simular, sovintejar, subministrar, sumar, tibar, tombar, transferir, transmetre, travessar, venerar, ventar, verificar.
2. **VASE** (39, 19,5%): abatre, acomiadar, acontentar, admirar, afegir, aferrar, ajeure, ajustar, apressar, aprofitar, avenir, barallar, beneficiar, col.locar, compadir, concretar, decantar, deduir, diferenciar, diluir, dipositar, disputar, dividir, encaminar, enfonsar, enllaçar, envoltar, estendre, ficar, filtrar, incorporar, lamentar, perllongar, precipitar, propagar, remuntar, separar, servir, trobar.
3. **Intransitius** (32, 16%): abellir, accedir, acostumar, acudir, agradar, al.ludir, brillar, caure, col.laborar, comparèixer, concordar, costar, créixer, descendir, dormir, emergir, esclatar, esmorzar, evolucionar, figurar, gaudir, nedar, néixer, ocórrer, oscil.lar, plaure, regalimar, somriure, sortir, sospirar, surar, trucar.

Figura 4.1: *Classificació manual de 200 verbs*

4.2.2 Definició de trets

Per extreure dades que puguin definir els verbs i puguin ser emprades per l'algoritme per classificar-los, hem emprat els mateixos trets que ja havíem fet servir pels experiments amb mètode supervisat (vegeu la secció 3.3): DirClitic, DetONom, Passiva, Punt, Prep, Se, DetAmbNom, NoConcordança, VerbNoPersonal i V12.

Podem veure la mitjana dels valors dels trets per a cada una de les nostres tres classes a la taula 4.1.

Tret	Transitiu	VASE	Intransitiu
DirClitic	4,8	4,6	0,5
DetoNom	26,4	16,3	14,1
Passiva	6,5	3,1	0,6
Punt	7,1	6,8	10,9
Prep	17,3	31,3	40,2
Se	11,8	33,8	2,6
DetAmbNom	31,9	27,6	23
NoConcordança	28,4	26,5	13,2
VerbNoPersonal	54,6	41,7	18,6
V12	12,4	12,2	3,1

Table 4.1: Valors dels trets per als transitius, VASE i intransitius

Els verbs transitius tenen valors molt alts pels trets DirClitic, DetoNom, Passiva, DetAmbNom, NoConcordança, VerbNoPersonal i V12; tenen valors baixos pels trets Punt i Prep i uns valors mitjans (entre VASE i intransitius) per Se. Els verbs VASE tenen valors alts per DirClitic i V12 (com els transitius) i Se (a diferència dels transitius); valors mitjans per DetONom, Passiva, Prep, NoConcordança, VerbNoPersonal i valors baixos per Punt. Finalment, els verbs intransitius tenen valors baixos per a tots els trets excepte Punt i Prep.

4.2.3 Paràmetres experimentals

Cluto ofereix diverses opcions per dur a terme els experiments. Les opcions que vàrem triar corresponen a les de l'algoritme anomenat *K-means* (Zhao i Karypis, 2001). En aquest algoritme, cada *cluster* està representat pel seu centroid. El centroid d'un *cluster* és un vector format per un conjunt de trets. El valor de cada tret del centroid és igual a la mitjana dels valors que els objectes del *cluster* tenen per a aquell tret. Per tant, seria la representació de l'objecte prototípic del *cluster*, que pot coincidir amb un objecte real o no.

L'algoritme agrupa els objectes de manera que els objectes d'un *cluster* siguin al màxim de semblants al centroid d'aquest *cluster*. Al principi, l'algoritme fa els *clusters* a l'atzar i en

calcula els centroides. A continuació, canvia de *cluster* aquells objectes que es troben més a prop del centre d'un altre *cluster*, de manera que tots els objectes estiguin en el *cluster* el centre del qual està més a prop. L'algoritme torna a calcular els centroides dels nous *clusters* i mou altre cop els objectes que estan més a prop del centre d'un altre *cluster* que no pas del centre del *cluster* en què es troben. Aquest procés es repeteix el nombre de vegades que l'usuari indiqui.

Els paràmetres concrets que vàrem especificar són els següents:

- Mètode de *clustering*. Hi ha diversos mètodes per fer l'agrupament d'objectes. En el mètode directe, la solució de *clustering* es computa buscant simultàniament tots els *clusters* que es volen (l'usuari ha d'indicar el nombre de *clusters*). Aquest mètode s'oposa al de bisecció repetida, en què, en primer lloc, es fan dos grups i després es torna a dividir un dels *clusters* fins arribar al nombre de grups desitjat. En general, aquest segon mètode sol anar millor si els objectes estan descrits per molts trets. En canvi, si el nombre és relativament petit (menys de 20), el mètode directe és més eficient. Com que hem estat treballant amb 10 trets s'ha triat el mètode directe. Tant el mètode directe com el de bisecció repetida són parcials, en contraposició als mètodes aglomeratius (en què els objectes es van agrupant en parelles fins obtenir el nombre de *clusters* desitjat).
- Funció de similitud. S'ha mesurat la similitud entre els objectes fent servir el cosinus, mesura estàndard per a aquesta mena d'experiments.
- Criteri de *clustering*. Hi ha dos grans tipus de criteris que es poden emprar per trobar els *clusters*: els criteris interns (que intenten que els objectes d'un mateix *cluster* s'assemblin al màxim possible entre ells) o els externs (que intenten que cada *cluster* sigui al més diferent possible dels altres *clusters*). Nosaltres hem emprat un criteri intern, que s'anomena \mathcal{I}_2 , la fórmula del qual es troba a 4.1¹.

$$\text{maximitza } \mathcal{I}_2 = \sum_{r=1}^k \sum_{d_i \in S_r} \cos(d_i, C_r) \quad (4.1)$$

K és el nombre total de *clusters*, S el conjunt de verbs amb el qual farem el *clustering*, C_r el centre del *cluster* r i d_i el vector d'un dels objectes. Per a cada *cluster*, l'algoritme maximitza la similitud entre el cosinus de cada verb i el cosinus del centre del *cluster*. Per tant, es pretén que cada objecte sigui al més semblant possible al centre del *cluster* al qual pertany.

¹També s'ha provat amb altres criteris i els resultats eren semblants.

- Nombre de *clusters*: S'han fet proves amb diversos nombres de *clusters*: de 3 a 7 *clusters*. Tanmateix, com que hem establert tres classes, ens centrarem en la solució en 3 *clusters*. La solució en 4 *clusters* també serà útil per fer algunes observacions.
- Nombre de solucions de *clustering*. Com que, per començar l'experiment, els *clusters* es fan a l'atzar, podria passar que el resultat depengués massa dels *clusters* amb els quals s'ha començat l'experiment. Per tal d'evitar-ho, l'algoritme computa diverses solucions de *clustering* emprant diferents *clusters* inicials i escull la que ofereix millors resultats segons el criteri de *clustering*. El nombre de solucions computades va ser 30.
- Nombre d'iteracions. És el nombre de vegades que es fan variacions, que es canvien verbs de *clusters* per tal de maximitzar els resultats segons el criteri de *clustering*. En els nostres experiments, el nombre d'iteracions és 20.

Com a output, Cluto (Karypis, 2002) ofereix diversa informació:

- La llista de verbs que conté cada *cluster*.
- Els trets més descriptius de cada *cluster* i els trets que més discriminen cada *cluster* de la resta. Un tret és descriptiu si els seus valors tenen valors semblants per a tots els objectes del mateix *cluster*. Un tret és discriminant si contribueix a diferenciar un *cluster* dels altres. Moltes vegades, els mateixos trets són els descriptius i discriminants d'un *cluster*.
- Un gràfic que permet veure com estan relacionats els *clusters* entre ells i quina mena de valors té el centroide de cada *cluster* per a cada tret (la figura 4.3 n'és un exemple).

4.3 Resultats

4.3.1 Solució en tres *clusters*

Hem dut a terme els experiments de *clustering* de dues maneres diferents: fent el *clustering* només amb els 200 verbs que teníem anotats i fent-lo amb els 1288 verbs que apareixen al corpus més de 50 vegades². La distribució dels 200 verbs en tres *clusters* per cadascuna de les opcions es pot observar a la figura 4.2. Cada columna representa un *cluster* i cada color una classe: de més clar a més fosc, transitius, VASE i intransitius.

Si fem els experiments només amb 200 verbs, trobem que el *cluster* 0 inclou la gran majoria de verbs transitius (115) i alguns intransitius (5) i VASE (7). El *cluster* 1 inclou la majoria

²Dels 1291 que apareixen més de 50 vegades, hem eliminat els tres verbs atributius: *esdevenir*, *semblar*, *estar*.

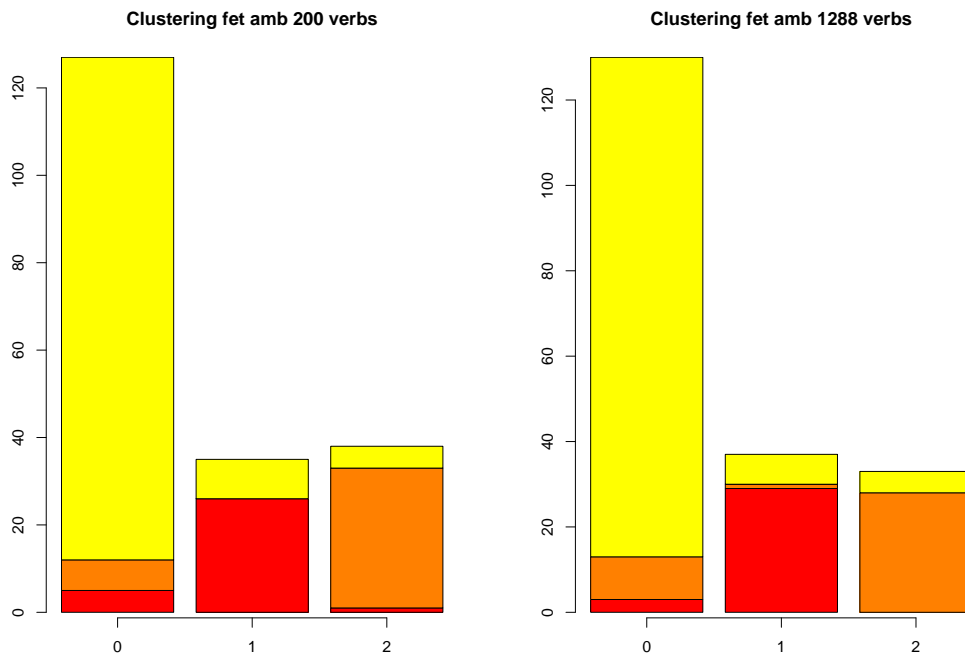


Figura 4.2: Solució en 3 clusters. Els colors representen, de més clar a més fosc, transitius, VASE i intransitius

d'intransitius (26) i alguns transitius (9). Finalment, el *cluster* 2 inclou la majoria de VASE (32), alguns transitius (5) i 1 intransitiu. Si fem els experiments amb 1288 verbs, la distribució és força semblant. Els transitius i els intransitius queden millor agrupats, mentre que els VASE queden una mica més repartits. El fet que la distribució de verbs en *clusters* presenti poques diferències tant si fem els experiments amb 200 verbs com si els fem amb 1288 és un positiu per demostrar la robustesa del sistema. Si aquest sistema funcionés només quan es fan servir els 200 verbs anotats manualment, probablement voldria dir que els trets s'han anat refinant pensant *ad hoc* en aquests verbs concrets i que, quan es volen tractar verbs diferents dels que s'han anotat manualment, el sistema deixa de funcionar. Un sistema que presentés aquest problema, evidentment, no seria útil per extreure informació de manera automàtica de verbs nous dels quals no es disposa informació.

A la taula 4.2, podem veure les dades de cobertura, precisió i el coeficient F per a cada *cluster*, mesures estàndard d'avaluació (Manning i Schuetze, 1999). La precisió és la proporció d'elements en un *cluster* que estan ben classificats i es calcula dividint el nombre de verbs ben classificats entre el nombre total d'elements en el *cluster*. És a dir, una precisió alta indica que hi ha pocs elements mal classificats en el *cluster*. La cobertura és la proporció d'elements ben classificats que estan en el mateix *cluster* i es calcula dividint el nombre d'elements ben classificats que hi ha en un *cluster* entre el nombre total d'elements d'aquella classe. Una

cobertura alta indica que la majoria d'elements d'una mateixa classe estan en el mateix *cluster*. El coeficient F combina la cobertura i la precisió en una sola mesura d'avaluació. Si volem donar el mateix pes a la precisió i a la cobertura, tal com fem aquí, simplement cal fer la mitjana d'aquests dos valors.

<i>Cluster</i>	Precisió	Cobertura	Coefficient F
0 Transitius	0,90	0,89	0,895
1 Intransitius	0,74	0,81	0,775
2 VASE	0,84	0,82	0,83

Table 4.2: Índexs de cobertura, precisió i coeficient F

Com hem explicat a la secció 4.2.3, Cluto ofereix diversa informació sobre la manera com l'algoritme fa l'agrupament en *clusters*. El *cluster* 0 és sempre el més compacte i l'últim (el 2, en aquest cas) el menys compacte. En aquest sentit, és significatiu, i esperable, que el *cluster* 0 correspongui als transitius (la classe, de llarg, més nombrosa) i que el *cluster* 2 correspongui als VASE (classe més inestable i intermitja entre les altres dues).

Els trets més descriptius per a cada *cluster* segons l'algoritme són:³

- *Cluster* 0: VerbNoPersonal (50,4%), DetAmbNom (14,5%), NoConcordança (13,5%), DetONom (11,1%), Prep (3,8%).
- *Cluster* 1: Prep (59,7%), DetAmbNom (18%), DetONom (7%), VerbNoPersonal (5,7%), Punt (5,4%).
- *Cluster* 2: VerbNoPersonal (24,3%), Se (23,3%), Prep (17,5%), DetAmbNom (15,1%), NoConcordança (12,4%).

El percentatge després de cada tret és el percentatge d'acord per a aquest tret dels objectes del *cluster*. El *cluster* dels transitius (el 0) és definit, principalment, per VerbNoPersonal, un tret destinat a marcar transitivitat i, en menor mesura, per altres trets que també pretenen descriure els transitius. També hi apareix el tret Prep, però amb un percentatge d'acord molt baix. En canvi, com era esperable, aquest tret és el que més pes té per definir el *cluster* dels intransitius (*cluster* 1), classe que inclou els verbs amb complement de règim. L'altre tret que es va establir per detectar intransitivitat, el tret Punt, apareix amb un percentatge baix. Tanmateix, és significatiu que aquest tret només el trobem com a tret descriptiu en aquest *cluster* i no en

³Totes les dades que presentarem en aquesta secció i a l'apartat 4.3.3 fan referència a l'experiment de *clustering* fet només amb els 200 verbs anotats manualment. Els resultats dels experiments en què s'empraven 1288 verbs no presenten diferències significatives

cap altre. També trobem trets destinats a marcar transitivitat, però amb percentatges d'acord molt menors als del *cluster* 0.

El *cluster* dels VASE (el 2) no té un tret descriptiu principal que tingui un percentatge d'acord molt per sobre dels altres, cosa que sí que passava amb els altres *clusters*. Més aviat, sembla que aquest *cluster* està definit per un conjunt de trets que tenen un percentage d'acord semblant (amb un valor important, però més baix que el del tret més descriptiu dels *clusters* 0 i 2). Sembla que els trets VerbNoPersonal, DetAmbNom i NoCondordança contribueixen a identificar les alternances transitives d'aquests verbs. Tanmateix, es tracta d'una transitivitat menys marcada que la dels verbs del *cluster* 0, cosa que es veu especialment amb el tret VerbNoPersonal (50,4% en el *cluster* 0 i 24,3% en el *cluster* 2). Els trets Prep i Se, en canvi, identifiquen les alternances intransitives d'aquests verbs. El tret Prep té menys importància que en el *cluster* dels intransitius (17,5% versus 59,7%) i el tret Se només apareix com a tret descriptiu en aquest *cluster*. Aquest conjunt de trets descriptius del *cluster* 2 reforça la idea que la classe VASE és una classe intermitja entre les altres dues, que té aspectes en comú tant amb els transitius com amb els intransitius.

A la taula 4.3 podem veure les preposicions més freqüents que segueixen els verbs VASE. Com era d'esperar *a* i *de* són les més freqüents, seguides de *en* i *amb*.

Preposició	Num. verbs
a	83
de	50
en	45
amb	35
per	25
cap	3
sobre	3
fins	1

Taula 4.3: Preposicions més freqüents que segueixen els verbs VASE

La figura 4.3 mostra com estan agrupats els *clusters* que s'han format. Per construir aquesta agrupació, l'algoritme ajunta els dos *clusters*, de manera que la solució de *clustering* que en resulti sigui la millor possible segons el criteri escollit. Segons l'algoritme, el *cluster* 0 i el 2 estan més relacionats entre ells que amb el *cluster* 1. És a dir, la divisió principal és entre intransitius i la resta (transitius i VASE) i la segona divisió és entre transitius i VASE.

En aquesta figura també es representen els valors dels trets per al centroide de cada *cluster* (com més intens és el color, més elevat el valor). Es pot veure que els trets amb el color més intens són els trets descriptius de cada classe.

Finalment, és també interessant d'observar la manera com s'han agrupat els trets. D'una banda, tenim els trets Punt i Prep (és a dir, els destinats a detectar intransitius) i, de l'altra, tota la resta. Pel que fa a l'agrupament de la resta de trets, trobem el tret Se (important per detectar VASE) agrupat amb dos trets que tenen valors molt baixos per totes tres classes (V12 i DirCLitic). També estan agrupats els trets que tenen valors alts per a la classe de transitius: VerbNoPersonal, NoConcordança i DetAmbNom, i també el tret Passiva, que és l'únic que té un valor baix per les tres classes. A l'annex B, hi hem inclòs la solució completa de *clustering* que retorna CLUTO.

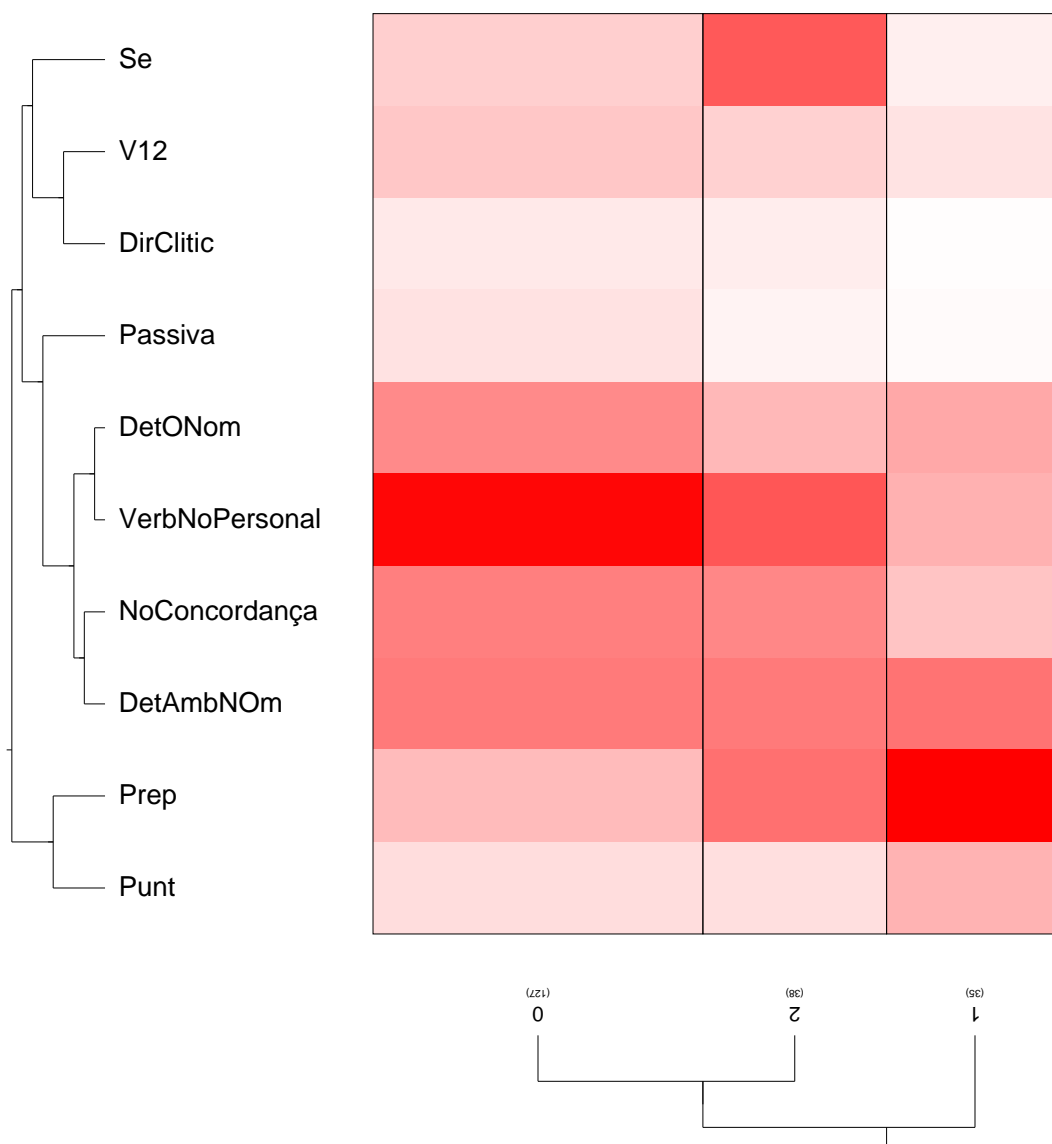


Figura 4.3: Característiques dels centroides dels 3 clusters

4.3.2 Anàlisi d'errors

En aquesta secció, revisarem els verbs que no han estat ben classificats. És a dir, analitzarem els verbs la classe dels quals no és la mateixa que la de la majoria dels verbs del *cluster* al qual han estat assignats. Ens centrarem en els verbs que han estat mal classificats quan el *clustering* s'ha fet amb 200 verbs.⁴

Per a l'avaluació dels nostres experiments, hem fet servir una classificació simbòlica i rígida: un verb és transitiu o bé no ho és; un verb és intransitiu o bé no ho és, etc. Tanmateix, sembla clar que hi ha verbs “més” transitius que d'altres. Aquesta gradualitat, però, no l'hem tinguda en compte, ja que tant pels humans com per les màquines és més fàcil treballar amb classificacions simbòliques. Per tant, dins el grup de verbs mal classificats, probablement hi haurà errors clars, però també hi haurà casos de més difícil classificació, verbs que es troben a la frontera entre dues classes.

A la taula 4.4, hi podem veure els verbs mal classificats en cada *cluster* i la classe a la qual pertànyen.

<i>Cluster</i>	Transitius	Intransitius	VASE
0 Transitius		agradar al.ludir esmorzar néixer regalimar	admirar afegir aprofitar compadir envoltar servir trobar
1 Intransitius	alterar cessar configurar consultar netejar operar pensar rectificar reposar		
2 VASE	avorrir coure errar espolsar intensificar	concordar	

Table 4.4: *Verbs mal classificats*

⁴Val a dir que no hi ha gaire diferències entre els errors que es produeixen amb els dos mètodes: hi ha vuit verbs mal classificats quan fem el *clustering* amb 200 verbs que han estat ben classificats amb 1288 verbs; i hi ha uns altres set verbs mal classificats quan fem el *clustering* amb 1288 i ben classificats amb 200 verbs.

- Dins el *cluster* dels verbs transitius, hi ha 5 verbs intransitius i 7 verbs VASE. Dos dels intransitius, *esmorzar* i *regalimar*, tenen alguns usos transitius:

- (44) a. Després vam esmorzar pa amb formatge i olives negres
 b. I el resultat, un espectacle suggestiu que regalimava picardia, intel·ligència, ironia, imaginació

A més, *regalimar* i *néixer* són verbs amb una inacusativat molt marcada i, molt sovint, el seu subjecte apareix en posició postverbal, cosa que també succeeix amb *agradar*.

Els verbs VASE inclosos en aquest *cluster* probablement han estat mal classificats perquè en el corpus hi ha poques ocurrencies de l'alternança intransitiva (*verb incrementat amb se + preposició*). Sis dels set verbs VASE mal classificats en aquest *cluster* són els sis verbs de la classe VASE amb menys ocurrencies en aquesta alternança: entre el 2,04% (d'*envoltar*) i 9,09% (de *compadir*) de les ocurrencies del corpus. El verb *trobar* té l'onzè valor més baix (14,8%). Entre *compadir* i *trobar* hi ha quatre verbs VASE que han estat ben classificats al *cluster 2*. A la taula 4.5 trobem el percentatge d'ocurrencies intransitives dels verbs mal classificats i d'alguns verbs VASE ben classificats escollits a l'atzar. Podem veure que, en general, els verbs ben classificats presenten percentatges significativament més alts. Per tant, els errors de classificació es deuen a la manca d'ocurrencies de l'alternança intransitiva en el corpus.

Verbs mal classificats	% oc. intransitives	Verbs ben classificats	% oc. intransitives
admirar	3,12	beneficiar	30,65
afegir	4,36	diferenciar	18,28
aprofitar	6,95	encaminar	39,68
compadir	9,09	incorporar	25,30
envoltar	2,04	lamentar	13,14
servir	5,15	precipitar	28,71
trobar	14,8	propagar	28,14

Taula 4.5: Ocurrencies intransitives de verbs de la classe VASE

- Dins el *cluster* dels verbs intransitius, hi ha 9 verbs transitius (cap dels VASE). Aquests 9 verbs tenen en comú que van seguits de de nom o determinanat en un percentatge menor que la majoria de transitius. Quatre d'aquests verbs (*cessar*, *operar*, *rectificar*, *reposar*) van seguits de preposició en la majoria de casos; les ocurrencies de *pensar* van seguides, en primer lloc, de conjunció i, en segon lloc, de preposició. Alguns d'aquests verbs sovint s'usen sense objecte (per exemple, *resposar* o *netejar*) o tenen una alternança que requereix un complement preposicional (*pensar en*, *cessar de*).
- Dins el *cluster* dels verbs VASE, hi ha 5 verbs transitius i 1 intransitiu. Tres d'aquests

verbs (*avorrir*, *espolsar* i *intensificar*) tenen molta tendència a anar incrementats per la partícula *se* (més d'un 30% de les ocurrences del corpus). *Errar* va incrementat per *se* en un percentatge menor, però també significatiu (14,2%) i, a més, va seguit de preposició en un 20% de les ocurrences. Aquesta podria ser també la raó per la qual el verb *coure* ha estat classificat en el grup dels VASE. Tot i que, va incrementat per *se* en poques ocasions (5,7%), va seguit de preposició en un 23,5%. L'únic verb intransitiu mal classificat (*concordar*) va seguit de preposició en un percentatge important, però potser no prou significatiu per trobar-se en el *cluster* 1.

En resum, els verbs que han estat mal classificats són verbs poc canònics amb un comportament diferent del prototípic de la seva classe. Per tant, tenen valors diferents de la mitjana en els trets més descriptius del *cluster* al qual haurien d'estar. És a dir, es troben molt lluny del centroide del *cluster* de la classe a la qual pertanyen i, en canvi, segurament es troben més a prop del centroide d'un altre *cluster*, per la qual cosa es classifiquen malament.

4.3.3 Solució en quatre *clusters*

En la figura 4.4 es pot veure què succeeix si, en comptes de tres, classifiquem els verbs en quatre grups⁵. Es manté un *cluster* (el 0) que conté la gran majoria d'intransitius (26) i també 9 transitius i 1 VASE i un altre *cluster* (el 2) que conté la majoria de VASE (29) i 4 transitius. En canvi, el *cluster* de transitius de la solució amb 3 *clusters* es divideix en dos *clusters*. El *cluster* 1 conté la majoria de transitius (71) i alguns intransitius (4) i VASE (6). El *cluster* 3 conté també molts transitius (45) i alguns intransitius (2) i VASE (4)⁶.

És a dir, en la solució amb quatre *clusters*, se segueixen distingint correctament els transitius, intransitius i VASE. L'única diferència és que els transitius es divideixen en dos grups. Sembla que l'algoritme no fa cap divisió lingüísticament interessant dins els transitius. Observant els verbs dels dos *clusters* de transitius (1 i 3), es fa difícil veure quin criteri s'ha fet servir per separar-los. A la taula 4.6, es poden veure alguns verbs dels *cluster* 1 i 3 escollits a l'atzar.

Els trets descriptius i discriminants d'aquests dos *clusters* són els següents:

- *Cluster* 1
 - Descriptiu: VerbNoPersonal 78,7%, DetONom 14,5%, Preposició 1,5%.

⁵Aquesta classificació en quatre grups ens serà de molta utilitat en la secció 5.2.

⁶Totes aquestes dades es refereixen a l'experiment realitzat només amb 200 verbs. Com passava amb la solució amb 3 *clusters*, no hi ha grans diferències si l'experiment es fa amb 1288.

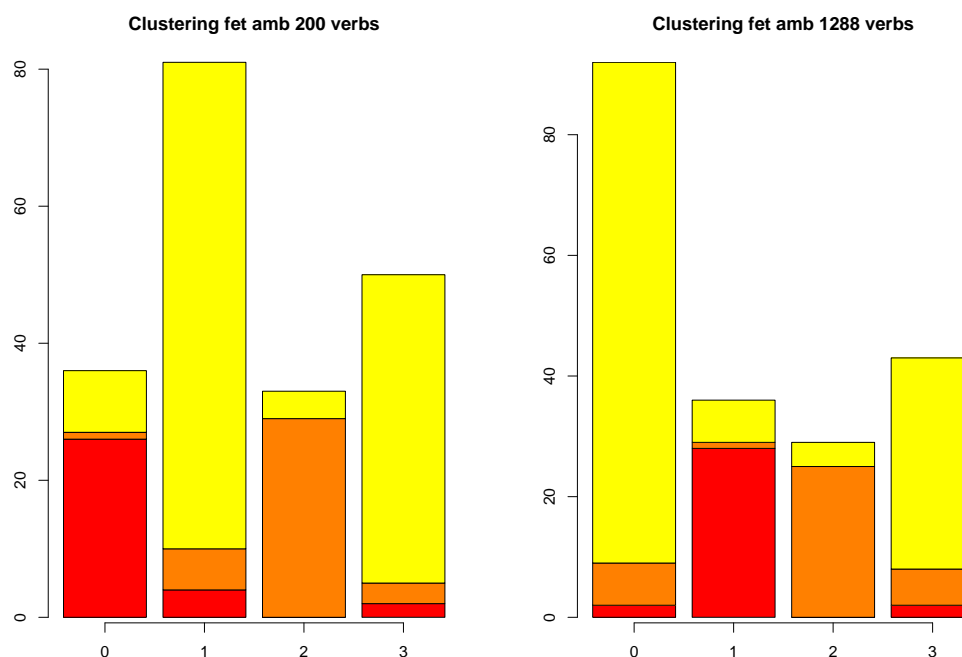


Figura 4.4: Solució en 4 clusters. Els colors representen, de més clar a més fosc, transitius, VASE i intransitius

<i>Cluster 1</i>	<i>Cluster 3</i>
celebrar	conèixer
destruir	edificar
obeir	odiar
regular	protegir
sovintejar	transmetre

Taula 4.6: Exemple dels verbs dels clusters 1 i 3 de la solució en 4 clusters

– Discriminant: Preposició 41,4%, VerbNoPersonal 38,4%, DetAmbNom 5,2%.

- *Cluster 3*

– Descriptiu: VerbNoPersonal 38,8%, DetONom 34,7%, NoConcordança 7,1%.

– Discriminant: DetONom 33,7%, Preposició 32,4%, NoConcordança 11%.

Podem veure que els dos trets més descriptius són els mateixos en tots dos *clusters*, VerbNoPersonal i DetONom, tot i que canvia el percentatge d'acord: 78,7% del *cluster 1* vs. 38,8% *cluster 3* pel tret VerbNoPersonal, 14,5% del *cluster 1* vs 34,7% del *cluster 3* pel tret DetONom. De fet, no és sorprenent que l'algoritme no faci cap distinció lingüística interessant dins els transitius, ja que no s'havia definit prèviament cap tret destinat a extreure automàticament una classe concreta de transitius.

L'ordre dels *clusters* és també destacable. El *cluster* menys compacte segons l'algoritme (el *cluster* 3, l'últim que fa) correspon a un dels dos *clusters* de transitius. És a dir, segons l'algoritme, la subclasse de transitius que detecta en el *cluster* 3 és menys compacte que la classe VASE, detectada en el *cluster* 2. A més a més, quan els experiments es fan amb només 200 verbs, el *cluster* 0 no és un *cluster* de transitius, sinó el dels intransitius, classe que, en aquestes circumstàncies, guanya en robustesa a la dels transitius.

Com que havíem definit tres classes, però, alhora, demanàvem que l'algoritme fes quatre *clusters*, hauria pogut succeir que un dels *clusters* fos mixte; és a dir, contingués elements de cada classe a parts iguals. Tanmateix, tot i extraient quatre grups, cosa que fa l'experiment més difícil, cada *cluster* continua contenint majoritàriament elements d'una de les classes: el *cluster* 0 conté intransitius, el *cluster* 2 conté VASE i els transitius estan repartits en els *clusters* 1 i 3.

4.4 Conclusions

En aquest capítol, hem presentat els experiments de *clustering* de verbs. Els resultats són força satisfactoris, ja que, en el resultat amb tres *clusters*, cada *cluster* conté majoritàriament verbs d'una de les tres classes que havíem definit: transitius, intransitius i VASE. Per tant, disposem d'un sistema que ens permet classificar automàticament els verbs en aquestes tres classes. Si recordem els resultats de la taula 4.2, podem veure que tant la precisió com la cobertura són elevades. L'annex C inclou un llexicó format pels 1288 verbs emprats en els experiments classificats en les tres classes: transitiu, intransitiu i VASE (també inclou informació sobre els verbs que subcategoritzen un complement preposicional, com explicarem a la secció 5.3).

Els verbs que han estat mal classificats són aquells que mostren un comportament més allunyat del prototipus de la seva classe. Això es podria deure al fet que les ocurrències concretes d'aquests verbs que apareixen en el corpus fossin “poc canòniques” o “poc representatives”. Si fos així, els resultats millorarien si treballéssim només amb verbs que apareguin, per exemple, més de cent vegades, en comptes de més de cinquanta. Una altra opció seria disposar d'un corpus més gran, possibilitat que es podria estudiar en futurs treballs.

Una altra via per millorar seria fer una classificació i establir uns trets que ens permetessin fer distincions més fines i captar més matisos. Podríem establir altres classes, com, per exemple, la dels verbs *object-drop* (com *menjar*), la dels verbs transitius que presenten una alternança amb complement preposicional (com *pensar*), etc. Aquesta opció, tanmateix, presenta dificultats importants. D'una banda, caldria investigar si hi ha trets que es puguin extreure automàticament dels recursos dels quals disposem que puguin diferenciar les noves

classes que s'estableixin. D'altra banda, caldrà assegurar-nos que cada classe estigui prou ben representada en els verbs anotats manualment i en les ocurrències del corpus. En aquest sentit, probablement serà insuficient anotar 200 verbs a l'atzar i caldrà garantir que hi hagi prou objectes de cada classe.

Un dels aspectes que caldria millorar és la definició d'alguns trets, ja que sembla que alguns dels nostres trets, tot i ser rellevants, estan "poc aprofitats". Podem afirmar que tots els trets són rellevants, ja que vàrem dur a terme diversos experiments eliminant cada vegada un tret i la distribució que en sortia presentava índexs de precisió i cobertura més baixos. Tanmateix, alguns trets sembla que es fan servir poc. Per exemple, els trets V12 o DirCLitic no són trets descriptius o discriminants de cap *cluster*. A més, aquests dos trets, juntament amb Passiva, tenen valors molts semblants per tots tres *clusters* (v. la figura 4.3).

Experiments addicionals

5.1 Introducció

En aquest capítol, presentarem dues direccions cap a les quals hem estès l'experiment principal d'aquest treball, presentat en el capítol 4. En primer lloc, a la secció 5.2 demostrarem que el sistema de *clustering* resulta adequat també amb un corpus etiquetat de forma exclusivament automàtica. En segon lloc, a la secció 5.3, presentem un altre sistema de *clustering* destinat a distingir els verbs intransitius purs dels verbs que regeixen un complement preposicional.

5.2 *Clustering* amb un corpus etiquetat automàticament

En aquesta secció volem demostrar que el nostre sistema de *clustering* no només es pot fer servir amb un corpus com el CTILC (etiquetat semiautomàticament i corregit manualment), sinó que també funciona amb un corpus etiquetat exclusivament de manera automàtica.

5.2.1 La *Constraint Grammar*

Provarem el sistema de classificació verbal amb el corpus CTILC (el mateix que ja havíem emprat) etiquetat de manera automàtica amb un analitzador superficial anomenat CatCG (Alsina et al., 2002). La CatCG ens ofereix informació semblant a la que ens oferia l' anotació de l'IEC. És a dir, el lema de cada forma, la seva categoria morfològica i una etiqueta amb informació morfològica. A més, la CatCG també proporciona informació sintàctica, que, tanmateix, no farem servir per a l'extracció de dades. Podem veure un extracte del corpus etiquetat amb la CatCG a la figura 5.1¹.

Tot i que les etiquetes de la CatCG i de l'IEC codifiquen informacions semblants, són superficialment molt diferents, ja que segueixen dos etiquetaris diferents: l'IEC segueix l'etiquetari Eagles (1996) i la CatCG una versió diferent d'aquest mateix etiquetari (Morel et al., 1997).

¹Aquest extracte és el mateix que el que hem fet servir per il·lustrar l'anotació del CTILC (figura 3.2)

mot	lema	categoria	eti. morf.	eti. sint.
la	la_lo	Nom_Pron	N5-MS_REEC3FS	<P_Atr_CD_CD-clt_Subj
veié	veure	Verb	VDP3S-	VPrin
amb	amb	Prep	P	<AN_Advl
extrema	extrem	Adj	JQ-FS	<AN_AN>
claredat	claredat	Nom	N5-FS	<P
i	i	Conj	CC	Conj
molt	molt	Adv	D4	AA/A>
precisa	precisar_precís	Adj_Verb	JQ-FS_V=R=S-	<AN_AN>
en	en	Prep	P	<AN_Advl
la	el	Det	EA-FS	DN>
seva	seu	Adj	JP63FS	AN>
imatge	imatge	Nom	N5-FS	<P
.	.	PT	.	PT

Figura 5.1: Extracte del CIEC etiquetat amb la CatCG

En conseqüència, per tal d'emprar el nostre sistema amb el corpus etiquetat per la CatCG, hem hagut d'adaptar l'extracció de dades de tal manera que es puguin reconèixer les noves etiquetes. A més, també s'han tingut en compte les diferències teòriques que hi ha entre l'etiquetatge de l'IEC i el de la CG (Quixal, 2003). Per exemple, els participis que complementen un nom, com en *els al.ludits anglesos*, són considerats com a adjectius per l'IEC i com a verbs per la CatCG. A més, molts mots tradicionalment considerats com a pronoms són etiquetats per la CatCG com a determinants. L'etiquetari de l'IEC i el de la CatCG també segueixen criteris diferents per establir el lema d'una forma en participi. Per l'etiquetari de l'IEC, el lema de, per exemple, *cantada* és el participi masculí singular (*cantat*), mentre que per l'etiquetari de la CatCG és l'infinitiu (*cantar*). Això fa que, per un mateix verb, amb l' anotació de la CatCG es tinguin en compte més ocurrences que amb l' anotació de l'IEC i això fa que es produeixin algunes variacions en els valors dels trets. Aquesta diferència de criteri tindrà efectes importants en la detecció d'alguna de les classes, com veurem en la secció 5.2.2.

La taxa d'error de la CatCG està calculada en un 1,42%, un percentatge baix que sembla que permet dur a terme experiments com els que proposem en aquest treball. Hi ha certs casos en què la CatCG no pot decidir l'etiqueta morfològica d'una forma, per la qual cosa resta ambigua. Per exemple, a la figura 5.1, hi ha dos mots la categoria morfològica dels quals queda ambigua: el mot *la* és ambigu entre nom i pronom, i el mot *precisa* entre adjectiu i verb (les possibles categories i les possibles etiquetes estan unides per un guió baix). L'ambigüetat es troba entre un 7% i un 8%. Per a la nostra extracció de dades, només farem servir formes que hagin estat completament desambiguades.

Podem veure la mitjana dels valors dels trets per a cada una de les nostres tres classes

a la figura 5.1. Si es comparen aquestes mitjanes amb les mitjanes calculades amb les dades extretes amb l' anotació de l'IEC (taula 4.1), es pot veure que els valors són força semblants i que es guarden les proporcions que tenia cada tret per les tres classes: per exemple, VerbNoPersonal segueix tenint un valor molt alt per als transitius, mitjà per als VASE i baix per intransitius; Se té un valor alt per als VASE, mitjà per als transitius i baix per als intransitius, etc. DetAmbNom és l'únic tret que perd capacitat distintiva entre les classes. En el corpus etiquetat per l'IEC (CIEC d'ara endavant), presentava una mitjana alta per als transitius (31,9%), mitjana per als VASE (27,6%) i baixa per als intransitius (23%). En canvi, amb el corpus etiquetat amb la CatCG (CCG d'ara endavant) aquestes diferències, que tot i que no eren molt marcades sí que eren prou rellevants, es dilueixen molt més: transitius (38,5%), VASE (34,4%), intransitius (34,6%). Tanmateix, tot i que les diferències són molt minses, els intransitius segueixen mostrant valors més baixos que els transitius, que és el que esperaríem.

Tret	Transitius	VASE	Intransitius
DirClitic	5,1	4,5	0,3
DetoNom	23	14,2	13,8
Passiva	6,5	3	0,4
Punt	8,2	7,7	13,2
Prep	16,9	31,5	39,5
Se	10,4	28,9	1,7
DetAmbNom	38,5	34,4	34,6
NoConcordança	34,4	34,6	15,4
VerbNoPersonal	46,7	31,1	19,9
V12	9,5	8,6	4,9

Table 5.1: Valors dels trets amb el corpus etiquetat per la CatCG

5.2.2 Solució en tres clusters

A la figura 5.2, podem veure els resultats del *clustering* amb tres i quatre clusters emprant 200 verbs².

En la solució en 4 clusters, la distribució segueix sent la mateixa que amb el CIEC: el cluster 0 correspon als transitius, el cluster 1 als intransitius i el cluster 2 als VASE³.

- El cluster 0 conté 101 transitius, 3 intransitius i 5 VASE.
- El cluster 1 conté 27 intransitius, 2 VASE i 6 transitius.

²Els resultats dels experiments realitzats amb 1288 eren semblants als realitzats amb 200 verbs.

³Es va fer una prova sense emprar el tret DetAmbNom i els resultats empitjoraven lleugerament. Cal suposar que, tot i que les diferències de la mitjanes de les tres classes per a aquest valor eren petites, aquest tret contribueix a reforçar l'estructura dels objectes que altres trets estableixen de manera més clara.

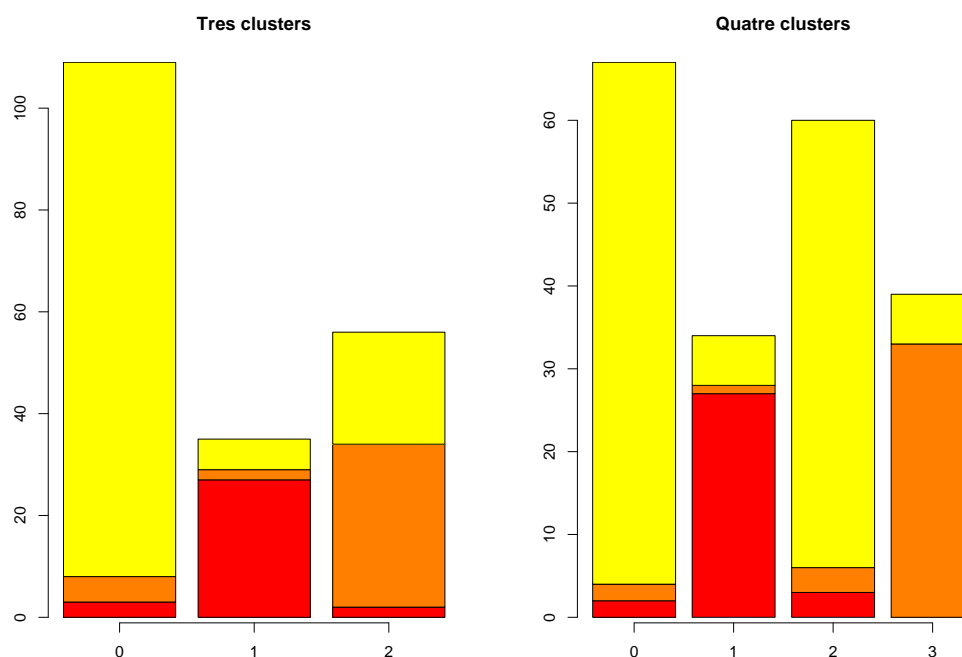


Figura 5.2: Solució en 3 i 4 clusters amb el CCG. Els colors representen, de més clar a més fosc, transitius, VASE i intransitius

- El *cluster* 2 conté 32 VASE, 22 transitius i 3 intransitius.

A la taula 5.2, podem veure els índexs de cobertura, precisió i coeficient F emprant el CCG.

<i>Cluster</i>	Precisió	Cobertura	Coefficient F
0 Transitius	0,92	0,78	0,85
1 Intransitius	0,77	0,84	0,80
2 VASE	0,84	0,57	0,70

Table 5.2: Índexs de cobertura, precisió i coeficient F amb el CCG

La detecció d'intransitius no es veu afectada negativament pel canvi de l' anotació humana a l' anotació automàtica, sinó que fins i tot, tant la precisió com la cobertura milloren una mica. En el *cluster* dels transitius millora la precisió, però empitjora la cobertura. La precisió millora ja que conté menys verbs no transitius (de 12 a 8), i la cobertura empitjora ja que conté menys transitius (de 115 a 101). La classe que més es ressenteix d'aquest canvi de sistema d' anotació del corpus és la VASE. La precisió del *cluster* 2 es manté igual (32 verbs VASE, 0,82), però la cobertura empitjora d'una manera important. El nombre de verbs que no són de la classe VASE passa de 6 a 24 (de 0,84 a 0,57); d'aquests 24 verbs, 22 són verbs transitius.

És a dir, la distinció entre intransitius i la resta de verbs es manté, i hi segueix havent un *cluster* majoritari de transitius. En canvi, la distinció entre transitius i VASE es difumina: si

sabem que un verb pertany al *cluster* 2, no hi ha prou elements per saber es tracta d'un VASE o d'un transitiu.

La mala cobertura del *cluster* 2 es deu al fet que l'algorisme no atorga prou pes al tret Se. Això ho podem deduir comparant els trets descriptius del *cluster* 2 quan els experiments es fan amb el CIEC i quan es fan amb el CCG.

- CIEC: VerbNoPersonal (24,3%), Se (23,3%), Prep (17,5%), DetAmbNom (15,1%), NoConcordança (12,4%).
- CCG: NoConcordança (29,4%), DetAmbNom (25%), Prep (14,8%), VerbNoPersonal (14,0%), Se (7,9%).

Podem veure que el tret més important per detectar els VASE, el tret Se, baixa d'un percentatge d'acord del 23,3% a tan sols un 7,9%, cosa que provoca que molts verbs transitius siguin inclosos en aquest *cluster*. Amb el CCG, no es detecta prou bé aquest tret. Això també ho podem veure comparant les taules 5.1 i 4.1, que contenen les mitjanes dels valors dels trets per a cada classe en els corpus CIEC i CCG. En el CIEC, el 33,8% de les ocurrencies dels verbs VASE té el tret Se. En el CCG, aquesta mitjana baixa al 28,9%. Aquesta diferència de gairebé 5 punts, tot i ser significativa, en principi, no hauria d'afectar tant als resultats, ja que les proporcions dels valors entre les classes es mantenen. Tanmateix, hi ha una altra diferència molt significativa⁴. En el CIEC, el tret Se és el que té la segona mitjana més alta, només superat pel tret VerbNoPersonal (un 41,7%). En canvi, en el CCG, el tret Se és el cinquè amb la mitjana més alta: és a dir, hi ha quatre trets que tenen mitjanes més altes: NoConcordança, DetAmbNom, VerbNoPersonal i Prep.

Aquesta diferència dels valors del tret Se no es deu a la detecció del context lingüístic superficial, que es fa de la mateixa manera amb totes dues anotacions, sinó que es deu a la diferència de criteri pel que fa a l'anotació dels lemes dels participis que hem comentat a la secció 5.2.1. Per exemple, el verb *diluir* té un valor per al tret Se del 51,3% amb el corpus CIEC i tan sols del 28,2% amb el corpus CCG. Tanmateix, amb ambdues anotacions, es detecten correctament 58 ocurrencies d'aquest verb seguit o precedit de la partícula *se*. La diferència es troba en que, segons el corpus CIEC, hi ha 113 ocurrencies del lema *diluir*, mentre que, segons el corpus CCG, n'hi ha 205, ja que també comptabilitza els participis. Com que els participis no van mai seguits o precedits de *se*, el percentatge amb el corpus CCG surt molt més baix i el tret Se perd poder distintiu. Aquest problema es podria solucionar en treballs futurs refinant

⁴Dono les gràcies a la Gemma Boleda per aquesta observació

l'extracció de dades.

5.2.3 Solució en quatre *clusters*

El problema de la baixa cobertura del *cluster* 2 es resol parcialment si, en comptes de fer 3 *clusters*, en fem 4. En aquest cas, com succeïa amb el CIEC, trobem un *cluster* majoritàriament d'intransitius, un altre de VASE i dos de transitius:

- El *cluster* 0 conté 63 transitius, 2 intransitius i 2 VASE. Té una precisió del 0,94 i una cobertura del 0,48.
- El *cluster* 1 (el dels intransitius) millora una mica la precisió (0,79, de 8 verbs no intransitius a 7) i manté la cobertura (0,84, 27 intransitius).
- El *cluster* 2 conté 54 transitius, 3 intransitius i 3 VASE. Té una precisió 0,90 i una cobertura del 0,41.
- El *cluster* 3 (el dels VASE) millora una mica la cobertura (0,84, de 32 a 33 VASE) i la precisió augmenta de manera molt important (0,84, de 24 verbs no VASE a 6). Per tant, un verb del *cluster* 3 té una probabilitat molt alta de ser un verb VASE, cosa que no passava en cap *cluster* de la solució amb 3 *clusters*.

És a dir, en la solució amb 4 *clusters*, trobem *clusters* que mostren uns nivells de precisió i cobertura força acceptables. De fet, la precisió i la cobertura són sensiblement millors que amb el corpus etiquetat manualment per a totes les classes, excepte per als transitius. Com que els transitius es troben dividits en dos *clusters*, aquests dos *clusters* (0 i 2) mostren una cobertura baixa en tots dos grups. Tanmateix, si es calculen les mesures d'avaluació ajuntant els dos *clusters* de transitius, s'obté una precisió del 0,92 i una cobertura del 0,90, dades també millors que amb els resultats del CIEC (compareu la taula 5.3 amb la taula 4.2).

<i>Cluster</i>	Precisió	Cobertura	Coefficient F
0 i 2 Transitius	0,92	0,90	0,91
1 Intransitius	0,79	0,84	0,81
3 VASE	0,84	0,84	0,84

Table 5.3: Índexs de cobertura, precisió i coeficient *F* amb quatre *clusters*

Per tant, sembla que aquest sistema per classificar verbs funciona tant amb un corpus corregit manualment com amb un corpus anotat de forma totalment automàtica.

A la taula 5.4 hi ha exemples triats a l'atzar de verbs transitius, classificats en els *clusters* 0 i 2. Com passava amb els experiments fets amb el CIEC, no observem que els transitius estiguin dividits en els dos *clusters* seguint un criteri lingüísticament interessant.

<i>Cluster 0</i>	<i>Cluster 2</i>
administrar	admetre
enfocar	enviar
escoltar	espia
regirar	seleccionar
travessar	venerar

Taula 5.4: Exemple de verbs dels clusters 0 i 2 de la solució en 4 clusters amb el CCG

Fent quatre *clusters*, la cobertura del *cluster* 2 millora perquè el tret Se és un tret important d'un dels *clusters*, cosa que no passava en la solució en quatre *clusters*. En aquest cas, el percentatge d'acord del tret Se és major que en la solució en tres *clusters*: 18,3% vs. 7,9%. Aquests són els trets descriptius d'aquest *cluster*:

- VerbNoConcordança 20,3%, DetAmbNom 19,1%, Se 18,3%, Prep 18,2%, VerbNoPersonal 17,2%.

En el *clustering* fet amb el CIEC hi ha 27 verbs mal classificats; amb el *clustering* fet amb el CCG n'hi ha 23, dels quals 14 coincideixen amb verbs mal classificats amb el CIEC i els altres 9 són verbs que amb el CIEC es havien estat ben classificats. A la taula 5.5, trobem un resum dels verbs mal classificats amb cada anotació.

Verbs mal classificats amb ambdues anotacions	Verbs mal classificats només amb CCG	Verbs mal classificats només amb CIEC
admirar	ajeure	agradar
afegir	brillar	al.ludir
aprofitar	connectar	alterar
avorrir	contrastar	compadir
cessar	costar	configurar
concordar	detallar	consultar
coure	espolsar	espolsar
envoltar	realitzar	errar
esmorzar	transmetre	néixer
intensificar		netejar
operar		pensar
regalimar		rectificar
reposar		trobar
servir		

Taula 5.5: Verbs mal classificats amb les diverses anotacions

5.2.4 Conclusions

En aquesta secció, hem repetit els experiments de *clustering* del capítol 4 amb el corpus etiquetat exclusivament de forma automàtica. Els resultats són fins i tot una mica millors que amb el corpus revisat manualment, com es pot veure a la taula 5.3. Això és molt important, ja que vol dir que si mai disposem d'un corpus més gran, simplement caldrà anotar-lo automàticament i el podrem fer servir per al nostre sistema de *clustering*.

5.3 Cap a una adquisició dels complement preposicionals

5.3.1 Introducció

En el capítol anterior, hem presentat un sistema que ens permetia classificar els verbs en tres classes: transitius, intransitius i VASE. Tanmateix, dins el *cluster* dels intransitius, hi trobem dos tipus de verbs de naturalesa molt diversa: intransitius purs (verbs que no admeten cap tipus de complement) i verbs de règim verbal (verbs que necessiten un complement introduït per preposició). En aquesta secció, aplicarem el mètode de *clustering* per classificar els verbs segons si porten un complement de règim verbal o no.

Per poder avaluar els resultats, hem classificat manualment els 212 verbs que es troben en el *clusters* dels intransitius (*cluster* 1) en els experiments fets amb 1288 verbs. Les classes que hem emprat són les següents:

1. Verbs intransitius purs: verbs que no admeten cap complement: ni directe, ni preposicional: *créixer, brillar, subsistir, treballar*.
2. Verbs amb complement de règim verbal: verbs que regeixen un complement preposicional: *accedir, confiar, maldar, tendir*. En alguns casos, es pot prescindir d'aquest complement (*parlar, pactar, presumir*).
3. Verb transitius: Verbs que han estat mal classificats en el *cluster* dels intransitius, ja que, de fet, són transitius: *actuar, operar*.

Són especialment difícils de classificar els verbs que apareixen freqüentment seguits d'un element que semànticament indica un lloc, ja que no és fàcil decidir si aquest element és un complement del verb o un adjunt. Per exemple:

(45) a. quan s'arriba *al nostre país*

- b. han marxat *de Bilbao*
- c. li permet de planar *per l'aire*
- d. rodolà *per les escales*

Hem fet servir una prova que, tot i que és indirecta, és útil per decidir si els sintagmes que segueixen aquests verbs són adjunts o complements. Es tracta d'una prova que s'ha fet servir per determinar la telicitat d'un verb. Segons Vendler (1967), les situacions denotades pels verbs es poden dividir de la següent manera:

- No processives
 - Estats (*states*): -Tèlic. *La casa és blanca*
 - Assoliments (*achievements*): +Tèlic. *He trobat la clau*
- Processives
 - Activitats (*activities*): - Tèlic. *Camino pel parc*
 - Realitzacions (*accomplishments*): + Tèlic. *Escriu una carta*

Les situacions processives són aquelles que poden dividir-se en diferents, fases, mentre que, en les no processives, això no és possible, ja sigui perquè denoten estats o perquè denoten situacions puntuals (assoliments). Pel que fa a la telicitat, una situació és tèlica si té un punt final inherent, més enllà del qual no pot continuar, mentre que una situació atèlica no té un punt final inherent (Huddleston i Pullum, 2002). *Escriure cartes* denota una situació atèlica, que no té un final marcat, mentre que *escriure una carta* és una acció tèlica; quan s'hagi acabat l'acció d'escriure la carta, s'acabarà l'acció. Les situacions tèliques admeten modificadors de límit, mentre que les atèliques admeten modificadors duratius:

- (46) a. Escriure una carta (en deu minuts)
- b. Escriure una carta (durant deu minuts)

Considererem que si un verb intransitiu, seguit d'un sintagma preposicional, admet un modificador de límit (és a dir, si denota una situació tèlica), el significat d'aquest sintagma preposicional és inherent al verb i, per tant, el classificarem com a verb de règim verbal. En canvi, si un verb i un sintagma preposicional denoten una situació atèlica, considererem aquest verb com un verb intransitiu pur. Així, seguint aquest criteri, considerariem, per exemple, *marxar* i *arribar* com a verbs de règim i *planar* i *rodolar* com a verbs intransitius.

- (47) a. quan s'arriba al nostre país en tres hores
- b. * quan s'arriba al nostre país durant tres hores
- c. rodolà per les escales *en cinc minuts
- d. rodolà per les escales durant cinc minuts

Els 212 verbs classificats apareixen a la figura 5.3.

1. **Intransitiu pur** (96): abellir, abundar, aflorar, ajeure, assentir, avançar, bategar, botar, brillar, brollar, cagar, caldre, caminar, caure, cavalcar, circular, comparèixer, conduir, costar, créixer, dansar, davallar, desaparèixer, descansar, dinar, dormir, embarcar, emmalaltir, emmudir, encaixar, enraonar, envellir, esclatar, evolucionar, figurar, florir, funcionar, jeure, lliscar, manar, mentir, morir, murmurar, navegar, nedar, néixer, obrar, ocórrer, opinar, perdurar, persistir, petar, planar, plaure, plorar, ploure, prevaler, progressar, prosperar, reaccionar, reparèixer, reular, rellicar, renéixer, replicar, ressonar, restar, retrocedir, riure, rodolar, saltar, seure, sobresortir, sobreviure, sobtar, soler, somriure, sopar, sospirar, suar, subsistir, succeir, surar, tintar, transitar, treballar, tremolar, triomfar, vacil.lar, variar, vibrar, viure, volar, xerrar, xisclar, xocar.
2. **Règim verbal** (107): accedir, acostumar, acudir, ajudar, al.ludir, arribar, ascendir, aspirar, assistir, baixar, cessar, coexistir, coincidir, col.laborar, començar, competir, comptar, concordar, concórrer, confiar, confluir, connectar, consistir, constar, contrastar, contribuir, convergir, conversar, convidar, conviure, correspondre, cuitar, culminar, datar, dependre, derivar, descendir, desembocar, diferir, disposar, dotar, dubtar, eixir, emergir, emigrar, ensopegar, entrar, equivaler, fiar, fruit, fugir, gaudir, incidir, incitar, induir, influir, ingressar, insistir, intervenir, invitar, jugar, lluitar, maldar, marxar, meditar, menar, obligar, oscil.lar, parlar, participar, partir, pecar, penetrar, pensar, pertocar, pertànyer, prescindir, presumir, privar, procedir, provenir, pujar, qualificar, radicar, raure, recaure, reeixir, reflexionar, remetre, renegar, renunciar, repercutir, residir, respondre, retornar, romandre, sortir, tardar, telefonar, tendir, titllar, topar, tornar, trigar, trucar, venir, viatjar.
3. **Transitiu** (9): actuar, arrelar, durar, encomanar, intentar, operar, pressionar, reposar, trametre.

Figura 5.3: Classificació manual dels 212 verbs

5.3.2 Trets

Hem emprat quatre trets per extreure dades sobre els verbs:

PrepMésFreq1: Aquest tret detecta quina és la preposició més freqüent que segueix un determinat verb i en calcula el percentatge d'aparició en relació amb el total de preposicions. Per exemple, *evolucionar* (310 ocurrències) va seguit de preposició en 162 ocasions (52,2%) i *tardar* (260 ocurrències) en 131 ocasions, cosa que representa un percentatge molt semblant a

evolucionar: 52,4%. Per tant, tots dos verbs presenten un valor semblant per al tret Prep que hem emprat en els capítols 3 i 4. Tanmateix, hi ha una diferència molt important entre tots dos verbs. En el cas d'*evolucionar* la preposició més freqüent és *cap*, que, amb 29 ocurrences representa el 17,9% de les vegades que aquest verb va seguit de preposició. També hi ha altres preposicions que segueixen el verb *evolucionar* amb certa freqüència, com *en* (20 oc.), *a* (18 oc.), *de* (11 oc.) i *amb* (10 oc). En canvi, en el cas de *tardar*, la preposició més freqüent és *a* (72 ocurrences), que representa un 54,9% de les vegades que aquest verb va seguit de preposició. Les altres preposicions que apareixen en el corpus seguint el verb *tardar* són molt menys freqüents que *a*: *en* (7 ocurrences), *per* (2 oc.), *devers*, *prop*, *des*, *entre*, *de* (1 oc.).

És a dir, sembla clar que *tardar* exigeix clarament una preposició regida, amb la qual cosa el nombre d'ocurrences d'aquesta preposició és clarament superior al de les altres. En canvi, a *evolucionar* la dispersió és molt més gran, cosa que ens fa pensar que no ens trobem davant d'un verb que exigeix complement de règim verbal, sinó d'un verb que va acompanyat d'adjunts amb molta freqüència.

PrepMésFreq2: Aquest tret detecta quina és la segona preposició més freqüent que segueix un determinat verb i en calcula el percentatge d'aparició en relació amb el total de preposicions. Aquest tret pretén detectar com a verbs de règim verbal aquells que poden regir dues preposicions diferents. Així, encara que el tret PrepMésFreq1 no presenti un valor molt alt, el fet que PrepMésFreq2 presenti un valor destacat ens pot ajudar a classificar aquest verb com a verb de règim verbal.

Puntuació: Verb seguit d'algun dels següents signes de puntuació: punt, punt i coma, dos punts, interrogació o exclamació. Els verbs intransitius haurien de mostrar valors més alts per a aquest tret que els verbs de règim verbal.

DetONom: Verb seguit de determinant o nom. Aquest tret hauria de mostrar valors alts per als verbs inacusatius o intransitius amb subjecte postverbal

A la taula 5.6 podem veure la mitjana dels valors dels trets per a cada categoria. Com era de preveure, els verbs de règim verbal tenen mitjanes més altes per als trets PrepMésFreq1 i PrepMésFreq2 i els verbs intransitius per als trets Puntuació i DetONom. Els verbs transitius, que recordem que havien estat mal classificats en aquest *cluster*, tenen valors alts per a DetONom i valors mitjans per als altres trets.

Tret	Règim	Intransitius	Transitius
PrepMésFreq1	35,19	10,72	13,11
PrepMésFreq2	9,79	5,76	7,08
Puntuació	5,63	13,36	6,24
DetONom	12,94	17,23	21,37

Taula 5.6: *Valors dels trets per als verbs de règim verbal, intransitius i transitius*

5.3.3 Resultats

Hem dut novament a terme experiments amb el *software* CLUTO (Karypis, 2002) amb els paràmetres que hem explicat a l'apartat 4.2.3. La distribució dels 212 verbs es pot observar a la figura 5.4. Els colors representen, de més clar a més fosc, els transitius, règim verbal i intransitius purs.

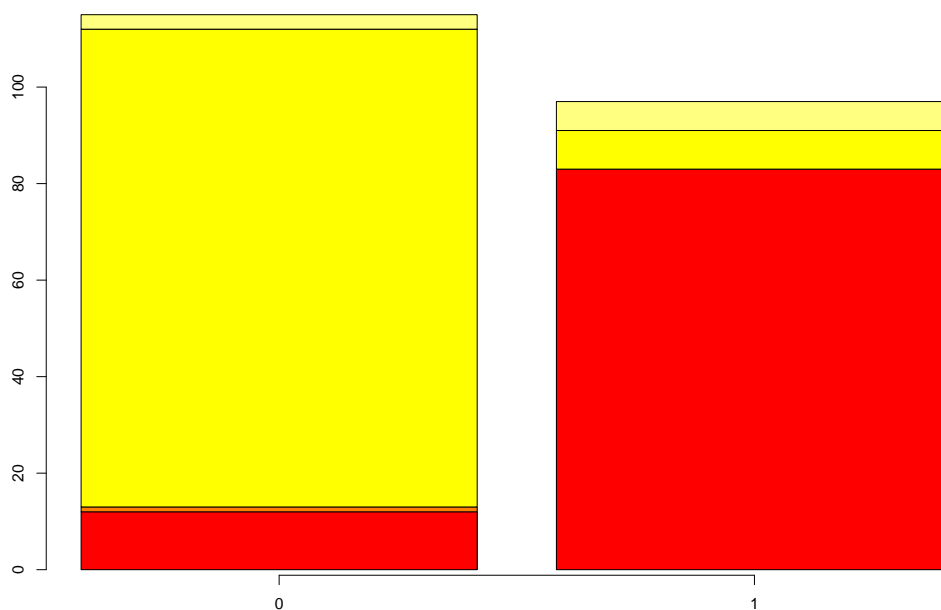


Figura 5.4: *Solució en 2 clusters. Els colors representen, de més clar a més fosc, transitius, règim verbal i intransitius purs*

El *cluster* 0 inclou 99 verbs de règim verbal, 13 intransitius i 3 transitius i el *cluster* 1 inclou 83 intransitius, 9 verbs de règim verbal i 6 transitius. A la taula 5.7 podem veure els índexs de cobertura, precisió i coeficient F.

Cluster	Precisió	Cobertura	Coefficient F
0 Règim	0,86	0,92	0,89
1 Intransitius	0,85	0,86	0,855

Taula 5.7: *Índexs de cobertura, precisió i coeficient F*

Els trets descriptius i discriminants de tots dos *clusters* són els següents:

- *Cluster 0*
 - Descriptius: PrepMésFreq1 77.0%, DetONom 12.5%, PrepMésFreq2 7.3%, Puntuació 3.3%.
 - Discriminant: PrepMésFreq1 49.1%, DetONom 26.5%, Puntuació 24.2%, PrepMésFreq2 0.2%.
- *Cluster 1*
 - Descriptius: DetONom 51.7%, Puntuació 28.3%, PrepMésFreq1 14.3%, PrepMésFreq2 5.7%.
 - Discriminant: PrepMésFreq1 49.1%, DetONom 26.5%, Puntuació 24.2%, PrepMésFreq2 0.2%.

El *cluster 0* està definit sobretot pel tret PrepMésFreq1 i el *cluster 1* pels trets DetONom i Puntuació. Sembla que el tret PrepMésFreq2 s'aprofita poc; tanmateix, presenta un percentatge d'acord major en el *cluster 0* que en el *cluster 1*. Per tant, les nostres previsions s'acompleixen i els trets serveixen per distingir la classe que havíem previst.

A la taula 5.8 veiem els verbs de règim que han estat classificats en el *cluster 1* (dels intransitius) i els verbs intransitius classificats en el *cluster 0* (dels verbs de règim verbal). Molts dels verbs intransitius mal classificats apareixen sovint amb un adjunt locatiu (*circular; conduir; transitar*) per la qual cosa tenen valors alts pels trets PrepMésFreq1 i, per tant, es troben al *cluster 0*. En canvi, els verbs de règim del *cluster 0* han estat mal classificats probablement perquè en el corpus sovint apareixen sense el complement preposicional (*jugat; protestar*) o perquè tenen alguns usos transitius (*baixar; pujar*).

Finalment, a la taula 5.9 incloem les preposicions més freqüents que segueixen els verbs que han estat classificats pel nostre sistema com a verbs de règim verb al. Com en el cas dels verbs VASE, les preposicions més freqüents són *a* i *de*, seguides de *en* i *amb*.

5.3.4 Conclusions

En aquesta secció hem presentat un sistema per distingir verbs intransitius i verbs de règim verbal. Com podem veure a la taula 5.7 els resultats són força correctes. Tanmateix, caldria seguir treballant en diverses direccions per millorar els resultats exposats aquí.

Verbs intransitius mal classificats	Verbs de règim mal classificats
abellir	baixar
ajeure	confluir
brollar	dotar
cagar	jugar
circular	marxar
conduir	meditar
costar	protestar
planar	pujar
reaccionar	sortir
rodolar	
tintar	
transitar	
xocar	

Taula 5.8: Verbs intransitius i de règim mal classificats

Preposició	Num. verbs
a	39
de	30
en	23
amb	14
per	6
damunt	1
entre	1
sobre	1

Taula 5.9: Preposicions més freqüents que segueixen els verbs del cluster 0

D'una banda, caldria millorar la classificació, establint més criteris que ens permetessin separar les classes i tractar millor els casos fronterers. D'altra banda, s'haurien d'establir més trets per dur a terme els experiments de *clustering*. Per a aquests experiments, ja hem obtingut resultats satisfactoris amb només 4 trets, però és de preveure que, si en definíssim més, els resultats del *clustering* es veurien reforçats.

Finalment, també es podria estendre l'adquisició de verbs amb complement preposicional a d'altres classes. Per exemple, es podria intentar extreure els verbs ditransitius (*regalar flors a la mare*) o els verbs que alternen entre un complement directe i un complement preposicional (*recórrer (contra) la sentència*).

A l'annex C, hi consta la informació de quina és la preposició més freqüent que segueix els verbs intransitius que han estat agrupats en el *cluster 0*, el *cluster* que agrupa la majoria dels verbs de règim verbal.

Conclusions i treball futur

En aquest treball, hem presentat diversos experiments destinats a classificar verbs automàticament segons la seva classe sintàctica. En aquest capítol, resumim els principals experiments i presentem les conclusions que en podem extreure, a més d'esbossar possibles direccions per ampliar aquest treball

6.1 Conclusions

En els capítols 3 i 4, hem presentat dues aproximacions diferents que compartien el mateix objectiu: disposar d'un sistema que fos capaç de classificar els verbs en diverses classes segons el seu comportament sintàctic. Per als experiments del capítol 3, hem emprat un mètode supervisat, que parteix d'una classificació prèvia i indueix regles per a detectar cada classe. En canvi, per als experiments del capítol 4, hem emprat un mètode no supervisat, el *clustering*, que no necessita informació prèvia sobre els objectes que cal classificar. En un primer moment, el mètode supervisat ens va semblar més adequat, ja que disposàvem d'una classificació manual i els resultats són més immediats i més fàcilment interpretables que no pas els dels mètodes no supervisats. Tanmateix, els resultats del mètode supervisat depenen enormement de la classificació prèvia, amb la qual cosa si aquesta no és prou coherent o no s'ajusta a les necessitats de l'experiment, els resultats estaran molt esbiaixats. Nosaltres ens vàrem trobar precisament amb aquest problema: la classificació no s'adaptava a les nostres necessitats, ja que no tenia en compte la freqüència d'ús dels verbs en cada marc de subcategorització per tal de classificar-los i, en canvi, per adquirir les classes hem fet servir informació bàsicament estadística.

El *clustering* no presenta aquests problemes. És un mètode més empíric que analitza els objectes a partir de les dades que els descriuen i, posteriorment, els agrupa. L'únic factor controlat per qui fa els experiments són els trets que han de servir per distingir les diferents classes. En el capítol 4 hem presentat un sistema de *clustering* que classifica els verbs en tres classes: transitius, intransitius i els verbs d'alternança amb *se* (VASE). Els resultats, tot i que contenen errors, són prou encoratjadors. La mitjana del coeficient F de la solució en tres *clusters* és d'un 83%, resultat comparable al dels treballs dels últims anys. Stevenson i Merlo

(2001), per exemple, arriben a un 69,8% d'encert en classificar verbs en *object-drop*, inergatius i inacusatius. Per tant, la nostra taxa d'encert és un 13% més elevada, si bé la seva tasca era més complexa que les que hem presentat aquí. Tanmateix, cal tenir en compte que el català té una sintaxi menys rígida que la de l'anglès, cosa que dificulta enormement l'extracció de dades.

Un dels aspectes més conflictius per l'adquisició de classes verbals és la gran heterogènia de patrons i d'alternances que trobem, especialment, dins els verbs transitius, així com la dificultat per classificar-los en subclasses discretes i ben definides. Com hem explicat en el capítol 4, hi ha verbs molt prototípics d'una classe i d'altres que no ho són tant, que es troben a la frontera entre dues classes. En aquests casos, un humà que classifiqui ha de prendre una decisió, que, en ocasions, estarà poc fonamentada o serà arbitrària. En aquest sentit, com que disposar d'una bona classificació humana és difícil, el *clustering* és una bona alternativa. Davant d'un verb poc canònic, un sistema de *clustering* també ha de prendre una decisió, però en aquest cas ho fa seguint les dades objectives que té sobre aquest verb i no està subjecte a la mena de prejudicis que tenim els humans.

Una de les aportacions d'aquest treball és la definició de deu trets que es poden extreure automàticament d'un corpus i que estableixen el grau de transitivitat (o de intransitivitat) d'un verb.

En la secció 5.3, hem aplicat novament la tècnica del *clustering* per dividir els intransitius en intransitius purs i verbs amb complement de règim verbal. Hem obtingut uns resultats bons amb un sistema que només feia servir quatre trets per descriure els verbs, ja que les classes definides presenten un correlat superficial molt clar i estable (presència o absència d'una determinada preposició seguint el verb).

A partir de la solució del capítol 3 en tres *clusters* i l'adquisició de verbs de règim verbal, hem creat el llexicó de l'annex C. Aquest llexicó consta de 1288 verbs classificats en transitius, intransitius i VASE i, en el cas dels intransitius de règim verbal i dels VASE, també hi consta la preposició que regeixen. Un llexicó d'aquestes característiques es pot emprar en gran diversitat de tasques de PLN: per exemple, per fer desambiguació morfològica o sintàctica.

En els nostres experiments, hem adquirit un nombre relativament petit de classes discretes. Això fa que no tots els verbs hi encaixin igual de bé, de tal manera que alguns són difícils de classificar. Tanmateix, un llexicó que segueixi aquesta mena de classificació és fàcilment "llegible" i manejable per un humà. A més, considerem que les classes adquirides són les més bàsiques, les que permeten de fer generalitzacions que afecten a un gran nombre de lemes i a partir de les quals es podrien seguir fent subdivisions més específiques.

6.2 Treball futur

Una possibilitat per continuar aquest treball seria ampliar el lexicó de l'annex C amb més informació sobre els verbs extreta automàticament. Per exemple, es podria adquirir informació sobre quins verbs poden anar complementats per una completiva. També es podrien definir subclasses de les nostres classes: per exemple, subtipus de transitius o verbs que participen en certes alternances, seguint la línia de treballs com els de Merlo i Stevenson (2001) o els de Lapata i Brew (2002). Tanmateix, aquesta és una opció complicada per diversos motius: d'una banda, per la complexitat de fer una bona classificació i, de l'altra, per la dificultat d'establir prous trets que distingeixin les diferents classes i que es puguin extreure automàticament amb els recursos que hem fet servir de moment. Probablement, emprar informació purament morfològica no seria suficient i caldria disposar d'informació de relacions de dependència.

També es podrien millorar i estendre els experiments presentats en la secció 5.3, destinats a adquirir automàticament verbs que regeixen complements preposicionals. D'una banda, caldria millorar els criteris per fer la classificació i especialment millorar la definició de verb de règim verbal. Per a aquests primers experiments, hem optat per emprar una prova que ens permetia de veure si el sintagma que seguia el verb era un complement inherent o un adjunt. D'altra banda, com ja hem dit en la secció 5.3, el sistema mitjançant el qual hem extret els verbs de règim verbal d'entre els intransitius es podria adaptar per tal de detectar els verbs transitius que tenen una alternança amb complement preposicional (per exemple, *creure (en) la mestra*). També es podria estendre el sistema per adquirir els verbs ditransitius o els que admeten complement indirecte. Tanmateix, això significaria modificar substancialment el mètode d'extracció de dades. Per detectar els trets emprats en tots els experiments que hem presentat, ens hem limitat a considerar la categoria morfològica de les paraules que segueixen o precedeixen el verb (amb una finestra de tres paraules per banda). El complement indirecte, però, se sol trobar més allunyat del verb i, per tant, no el podríem arribar a detectar. Probablement caldria disposar d'un corpus anotat sintàcticament, encara que fos de forma parcial, de manera que es pogués saber on comença cada sintagma.

Seria interessant també poder repetir els nostres experiments amb un corpus més gran, per tal de veure si els resultats milloren. Si això succeís, es demostraria que els verbs havien estat mal classificats en els primers experiments perquè en el corpus les ocurrències del verb no eren prou significatives o no exemplificaven prou bé la classe del verb. Si, en canvi, els resultats no milloressin, segurament voldria dir que el verb té un comportament especial, difícil de classificar. En aquest cas, caldria pensar en refinar o ampliar els trets per poder extreure dades més significatives.

Annex A. Regles extreteres fent servir la metodologia *cross-validation*

Aquí presentem el conjunt de regles extreteres pel programa Ripper fent servir la metodologia *cross-validation*.

Option: 10-fold cross-validation

----- run 1 -----

Hypothesis:

intr :- DirClitic \leq 0.00314785, Se \leq 0.00490196 (51/6).

intr :- DirClitic \leq 0, Se \leq 0.0672269 (43/15).

intr :- Passiva \leq 0.00675676, NoConcordança \leq 0.103896, Prep \geq 0.333333, Se \leq 0.309008, V12 \leq 0.0192308 (9/2).

default tr (914/6).

Error rate on holdout data is 7.75862%

Running average of error rate is 7.75862%

----- run 2 -----

Hypothesis:

intr :- DirClitic \leq 0.00314785, Se \leq 0.00847458 (60/15).

intr :- DirClitic \leq 0.00110988, Se \leq 0.0672269, NoConcordança \leq 0.0625 (22/5).

intr :- DirClitic \leq 0.00110988, VerbNoPersonal \leq 0.0416667, Se \leq 0.0875782 (6/1).

intr :- Passiva \leq 0.00178731, Se \leq 0.0153846, DetONom \leq 0.123249 (5/1).

default tr (920/11).

Error rate on holdout data is 3.44828%

Running average of error rate is 5.60345%

----- run 3 -----

Hypothesis:

intr :- DirClitic \leq 0.00110988, Se \leq 0.0672269 (92/29).

intr :- DetONom \leq 0.12766, Se \leq 0.00518135 (5/1).

default tr (907/12).

Error rate on holdout data is 3.44828%

Running average of error rate is 4.88506%

----- run 4 -----

Hypothesis:

intr :- DirClitic \leq 0.00222469, NoConcordança \leq 0.0454545, Se \leq 0.0875782 (60/9).

intr :- DirClitic \leq 0.00518135, Se \leq 0.00755668 (15/1).

intr :- DirClitic \leq 0.00518135, NoConcordança \leq 0.0963855, Se \leq 0.00755668 (12/4).

intr :- DirClitic \leq 0.0010929, Prep \geq 0.333333, Se \leq 0.309008 (14/5).

default tr (918/8).

Error rate on holdout data is 5.17241%

Running average of error rate is 4.9569%

----- run 5 -----

Hypothesis:

intr :- DirClitic \leq 0.00383632, Se \leq 0.0349727, NoConcordança \leq 0.0338983 (46/5).

intr :- DirClitic \leq 0.00518135, Se \leq 0.00565504 (30/6).

intr :- DirClitic \leq 0.00311526, Se \leq 0.0672269, Prep \geq 0.300699 (16/10).

intr :- NoConcordança \leq 0.0681818, DirClitic \leq 0, Se \leq 0.0578947 (5/1).

default tr (915/12).

Error rate on holdout data is 3.44828%

Running average of error rate is 4.65517%

----- run 6 -----

Hypothesis:

intr :- DirClitic \leq 0.00518135, Se \leq 0.00866739, NoConcordança \leq 0.0454545 (41/4).

intr :- DirClitic \leq 0.000128156, Se \leq 0.0578947 (53/21).

default tr (912/15).

Error rate on holdout data is 6.03448%

Running average of error rate is 4.88506%

----- run 7 -----

Hypothesis:

intr :- DirClitic \leq 0.00314785, Se \leq 0.044694 (48/6).

intr :- DirClitic \leq 0.000128156, Se \leq 0.0672269, NoConcordança \leq 0.0384615 (23/1).

intr :- DirClitic \leq 0.00383632, Prep \geq 0.300699, Se \leq 0.309008 (25/14).

intr :- DetONom \leq 0.0153453 (3/2).

intr :- Passiva \leq 0, Punt \geq 0.210084, Se \leq 0.0119048 (5/0).

default tr (913/6).

Error rate on holdout data is 6.89655%

Running average of error rate is 5.17241%

----- run 8 -----

Hypothesis:

intr :- DirClitic \leq 0.00314785, Se \leq 0.0672269 (97/37).

default tr (900/12).

Error rate on holdout data is 5.17241%

Running average of error rate is 5.17241%

----- run 9 -----

Hypothesis:

intr :- DirClitic \leq 0.00148368, Se \leq 0.00866739 (61/9).

intr :- DirClitic \leq 0.00518135, NoConcordança \leq 0.034965, Se \leq 0.0554017 (22/2).

intr :- DirClitic \leq 0.00383632, Prep \geq 0.300699, Punt \geq 0.0890411 (11/4).

default tr (920/17).

Error rate on holdout data is 6.03448%

Running average of error rate is 5.2682%

----- run 10 -----

Hypothesis:

intr :- DirClitic \leq 0.00383632, Se \leq 0.00755668 (65/13).

intr :- DirClitic \leq 0.00383632, NoConcordança \leq 0.0441176, Se \leq 0.0875782 (24/5).

intr :- DirClitic \leq 0.0010929, Se \leq 0.0349727, Prep \geq 0.300699, DetAmbNom \leq 0.111111 (10/1).

default tr (915/11).

Error rate on holdout data is 4.08475%

Running average of error rate is 4.55985%

===== statistical summary =====

Average error: 4.56% +/- 0.50%

Average time: 0.56 +/- 0.03

Annex B. Solució de *clustering*

En aquest annex, hi incloem un dels fitxers d'*output* de Cluto que resumeix la solució de *clustering* en tres *clusters* (dels experiments amb 200 verbs i amb el CIEC). Inclou la informació següent:

- *Matrix Information*: Nom del fitxer que conté la matriu de dades, nombre de files (objectes) i de columnes (trets) de la matriu.
- *Options*: Paràmetres triats per dur a terme l'experiment (coincideixen amb els explicats a l'apartat 4.2.3).
- *Solution*: Taula que conté per a cada *cluster*: el nombre d'elements, la similitud interna (ISim), la desviació estàndard d'ISim (ISdev), la similitud externa (ESim), la desviació estàndard d'ESim (ESdev), l'entropia (Entpy) i la puresa (Purty). A més, també inclou el nombre d'elements que hi ha de cada classe (transitius, intransitius i VASE).
- *Descriptive & Discriminating Features*: Llista dels trets descriptius i discriminants de cada *cluster* i el percentatge d'acord.
- *Hierarchical Tree that optimizes the I2 criterion function...*: Ordenació jeràrquica dels *clusters* en forma d'arbre i nombre d'elements de cada classe per a cada *cluster*.

Matrix Information

Name: /home/mayol/tesina/clustering/matriu, #Rows: 200, #Columns: 10

Options

CLMethod=Direct, CRfun=I2, SimFun=Cosine, #Clusters: 3 RowModel=None,
ColModel=None, GrModel=SY-DIR, NNbrs=40 Colprune=1.00, EdgePrune=-1.00,
VtxPrune=-1.00, MinComponent=5 CStype=Best, AggloFrom=0, AggloCRFun=I2,
NTrials=30, NIter=20

Solution

3-way clustering: [I2=1.85e+02] [200 of 200], Entropy: 0.396, Purity: 0.865

Cluster	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty	Tr	VASE	Intr
0	127	+0.881	+0.047	+0.669	+0.057	0.343	0.906	115	7	5
1	35	+0.785	+0.086	+0.583	+0.148	0.519	0.743	9	0	26
2	38	+0.848	+0.066	+0.736	+0.101	0.462	0.842	5	32	1

3-way clustering solution - Descriptive & Discriminating Features...

Cluster 0, Size: 127, ISim: 0,881, ESim: 0,669 Descriptive: VerbNoPersonal 50,4%, DetAmbNom 14,5%, NoConcordança 13,5%, DetONom 11,1%, Prep 3,8% Discriminating: Prep 54,3%, VerbNoPersonal 30,0%, Se 7,5%, DetONom 2,2%, NoConcordança 2,0%

Cluster 1, Size: 35, ISim: 0,785, ESim: 0,583 Descriptive: Prep 59,7%, DetAmbNom 18,0%, DetONom 7,0%, VerbNoPersonal 5,7%, Punt 5,4% Discriminating: Prep 48,4%, VerbNoPersonal 33,9%, NoConcordança 6,6%, Se 5,3%, Punt 3,1%

Cluster 2, Size: 38, ISim: 0,848, ESim: 0,736 Descriptive: VerbNoPersonal 24,3%, Se 23,3%, Prep 17,5%, DetAmbNom 15,1%, NoConcordança 12,4% Discriminating: Se 71,1%, VerbNoPersonal 13,4%, DetONom 9,3%, Prep 4,1%, Passiva 0,8%

Hierarchical Tree that optimizes the I2 criterion function...

4	Tr	VASE	Intr
!— 1	9	0	26
!- 3			
!— 2	5	32	1
!— 0	115	7	5

Annex C. Llexicó

En aquest annex, presentem 1288 verbs ordenats alfabèticament i classificats en tres tipus: transitius i intransitius i VASE, tal i com hem explicat en el capítol 4. També incloem informació sobre la preposició que segueix amb més freqüència els verbs VASE i els intransitius detectats com a verbs de règim verbal (vegeu el capítol 5.3). Els verbs intransitius purs estan classificats com a <intr> <0>. Hi ha 831 verbs transitius, 245 VASE i 212 intransitius, dels quals 97 són intransitius purs i 115 són verbs de règim verbal.

abaixar	<tr>		adaptar	<vase>	<a>
abandonar	<tr>		addicionar	<vase>	<a>
abastar	<tr>		aduir	<tr>	
abatre	<vase>	<sobre>	adequar	<vase>	<a>
abellir	<intr>	<de>	adherir	<vase>	<a>
abocar	<vase>	<a>	adir	<vase>	<amb>
abolir	<tr>		admetre	<tr>	
abonar	<tr>		administrar	<tr>	
abordar	<tr>		admirar	<tr>	
abraçar	<tr>		adobar	<tr>	
absorbir	<tr>		adoptar	<tr>	
abundar	<intr>	<0>	adorar	<tr>	
acabar	<tr>		adormir	<vase>	<en>
acariciar	<tr>		adornar	<tr>	
acaronar	<tr>		adquirir	<tr>	
accedir	<intr>	<a>	adreçar	<vase>	<a>
accelerar	<tr>		advertir	<tr>	
accentuar	<tr>		afaitar	<vase>	<a>
acceptar	<tr>		afanyar	<vase>	<a>
aclarir	<tr>		afavorir	<tr>	
aclucar	<tr>		afeblir	<tr>	
acollir	<tr>		afectar	<tr>	
acomiar	<vase>	<de>	afegir	<tr>	
acompanyar	<tr>		afermar	<vase>	<en>
acomplir	<tr>		aferrar	<vase>	<a>
aconseguir	<tr>		afinar	<tr>	
aconsellar	<tr>		afirmar	<tr>	
aconter	<vase>	<amb>	aflorar	<intr>	<0>
acordar	<vase>	<amb>	afluixar	<tr>	
acostar	<vase>	<a>	afrontar	<tr>	
acostumar	<intr>	<a>	agafar	<tr>	
acotar	<tr>		agitar	<vase>	<amb>
actuar	<intr>	<de>	agradar	<tr>	
acudir	<intr>	<a>	agrair	<tr>	
acumular	<vase>	<en>	agreujar	<vase>	<per>
acusar	<tr>		agrupar	<vase>	<en>

aguantar	<tr>	apagar	<tr>
aïllar	<tr>	apaivagar	<tr>
aixafar	<tr>	aparèixer	<tr>
aixecar	<tr>	apartar	<vase> <de>
ajeure	<intr> <a>	aplaudir	<tr>
ajornar	<tr>	aplegar	<tr>
ajudar	<intr> <a>	aplicar	<vase> <a>
ajuntar	<vase> <amb>	apoderar	<vase> <de>
ajupir	<vase> <per>	aportar	<tr>
ajustar	<vase> <a>	apreciar	<tr>
al.legar	<tr>	aprendre	<tr>
al.ludir	<intr> <a>	apressar	<vase> <a>
albirar	<tr>	aprofitar	<tr>
alçar	<tr>	aprofundir	<tr>
alegrar	<vase> <de>	apropar	<vase> <a>
alimentar	<vase> <de>	apropiar	<vase> <de>
allargar	<tr>	aprovar	<tr>
alliberar	<vase> <de>	aproximar	<vase> <a>
al·lotjar	<tr>	apuntar	<tr>
allunyar	<vase> <de>	argumentar	<tr>
alterar	<tr>	armar	<tr>
alternar	<tr>	arrabassar	<tr>
amagar	<vase> <en>	arranjar	<tr>
amenajar	<tr>	arrapar	<vase> <a>
amoïnar	<tr>	arreglar	<tr>
amollar	<tr>	arrelar	<intr> <en>
ampliar	<tr>	arrencar	<tr>
amuntegar	<vase> <en>	arreglerar	<vase> <a>
analitzar	<tr>	arreplegar	<tr>
anar	<vase> <a>	arribar	<intr> <a>
animar	<tr>	arriscar	<vase> <a>
anomenar	<tr>	arrodonir	<tr>
anotar	<tr>	arrosar	<tr>
anticipar	<vase> <a>	arrosegar	<tr>
anul·lar	<tr>	articular	<vase> <en>
anunciar	<tr>	ascendir	<intr> <a>

aspirar	<intr>	<a>	baixar	<intr>	<0>
assabentar	<vase>	<de>	ballar	<tr>	
assajar	<tr>		bandejar	<tr>	
assaltar	<tr>		banyar	<vase>	<a>
assassinar	<tr>		barallar	<vase>	<amb>
assecar	<vase>	<a>	barrar	<tr>	
assegurar	<tr>		barrejar	<vase>	<amb>
assemblar	<vase>	<a>	basar	<vase>	<en>
assentar	<vase>	<en>	bastar	<tr>	
assentir	<intr>	<0>	bastir	<tr>	
assenyalar	<tr>		bategar	<intr>	<0>
asseure	<vase>	<a>	batejar	<tr>	
assignar	<tr>		batre	<tr>	
assimilar	<tr>		bellugar	<vase>	<de>
assistir	<intr>	<a>	beneficiar	<vase>	<de>
associar	<vase>	<a>	beneir	<tr>	
assolir	<tr>		besar	<tr>	
assumir	<tr>		beure	<tr>	
atacar	<tr>		botar	<intr>	<0>
atansar	<vase>	<a>	brandar	<tr>	
atendre	<tr>		brillar	<intr>	<0>
atènyer	<tr>		brodar	<tr>	
atorgar	<tr>		brollar	<intr>	<de>
atrapar	<tr>		bufar	<tr>	
atreure	<tr>		buidar	<tr>	
atribuir	<vase>	<a>	bullir	<tr>	
aturar	<vase>	<a>	burxar	<tr>	
augmentar	<tr>		buscar	<tr>	
autoritzar	<tr>		cabre	<tr>	
avaluar	<tr>		caçar	<tr>	
avançar	<intr>	<0>	cagar	<intr>	<en>
avenir	<vase>	<a>	calar	<tr>	
avergonyir	<vase>	<de>	calcular	<tr>	
avisar	<tr>		caldre	<intr>	<0>
avorrir	<vase>	<de>	callar	<tr>	
badar	<tr>		calmar	<tr>	

caminar	<intr>	<0>	colpejar	<tr>	
canalitzar	<tr>		colpir	<tr>	
cansar	<vase>	<de>	comandar	<tr>	
cantar	<tr>		combatre	<tr>	
canviar	<tr>		combinar	<tr>	
capgirar	<tr>		començar	<intr>	<a>
captar	<tr>		comentar	<tr>	
capturar	<tr>		comercialitzar	<tr>	
caracteritzar	<vase>	<per>	cometre	<tr>	
caragolar	<vase>	<a>	commoure	<tr>	
carregar	<tr>		compadir	<tr>	
casar	<vase>	<amb>	comparar	<tr>	
castigar	<tr>		comparèixer	<intr>	<0>
caure	<intr>	<0>	compartir	<tr>	
causar	<tr>		compensar	<tr>	
cavalcar	<intr>	<0>	competir	<intr>	<amb>
cavar	<tr>		complaure	<vase>	<a>
cedir	<tr>		complementar	<vase>	<amb>
celebrar	<tr>		completar	<tr>	
centrar	<vase>	<en>	complicar	<vase>	<amb>
cercar	<tr>		complir	<tr>	
cessar	<intr>	<de>	compondre	<vase>	<de>
circular	<intr>	<per>	comportar	<tr>	
citar	<tr>		composar	<tr>	
clamar	<tr>		comprar	<tr>	
classificar	<tr>		comprendre	<tr>	
clavar	<tr>		comprometre	<vase>	<a>
cloure	<tr>		comprovar	<tr>	
cobrar	<tr>		comptar	<intr>	<amb>
cobrir	<tr>		comunicar	<tr>	
coexistir	<intr>	<amb>	concebre	<tr>	
coincidir	<intr>	<amb>	concedir	<tr>	
col.laborar	<intr>	<en>	concentrar	<vase>	<en>
col.locar	<tr>		concertar	<tr>	
colgar	<vase>	<amb>	concloure	<tr>	
collir	<tr>		concordar	<intr>	<amb>

concórrer	<intr>	<a>	contenir	<tr>	
concretar	<vase>	<en>	contestar	<tr>	
condemnar	<tr>		continuar	<tr>	
condicionar	<tr>		contractar	<tr>	
conduir	<intr>	<a>	contradir	<tr>	
conèixer	<tr>		contrastar	<intr>	<amb>
confeccionar	<tr>		contreure	<tr>	
confegir	<tr>		contribuir	<intr>	<a>
conferir	<tr>		controlar	<tr>	
confessar	<tr>		convèncer	<tr>	
confiar	<intr>	<en>	convenir	<tr>	
configurar	<tr>		convergir	<intr>	<en>
confirmar	<tr>		conversar	<intr>	<amb>
confluir	<intr>	<0>	convertir	<vase>	<en>
confondre	<vase>	<amb>	convidar	<intr>	<a>
conformar	<vase>	<amb>	conviure	<intr>	<amb>
connectar	<intr>	<amb>	convocar	<tr>	
conquerir	<tr>		coordinar	<tr>	
conrear	<tr>		copiar	<tr>	
consagrar	<tr>		copsar	<tr>	
consentir	<tr>		coronar	<tr>	
conservar	<tr>		corregir	<tr>	
considerar	<tr>		córrer	<tr>	
consignar	<tr>		correspondre	<intr>	<a>
consistir	<intr>	<en>	corroborar	<tr>	
consolar	<tr>		cosir	<tr>	
consolidar	<tr>		costar	<intr>	<de>
constar	<intr>	<de>	coure	<vase>	<a>
constatar	<tr>		crear	<tr>	
constituir	<tr>		créixer	<intr>	<0>
construir	<tr>		cremar	<tr>	
consultar	<tr>		creuar	<tr>	
consumir	<tr>		creure	<vase>	<en>
contaminar	<tr>		criar	<tr>	
contar	<tr>		cridar	<tr>	
contemplar	<tr>		crystal.litzar	<tr>	

criticar	<tr>		desaparèixer	<intr>	<0>
cuidar	<vase>	<de>	desar	<tr>	
cuinar	<vase>	<a>	desbordar	<tr>	
cuitar	<intr>	<a>	descansar	<intr>	<0>
culminar	<intr>	<en>	descarregar	<tr>	
cultivar	<tr>		descartar	<tr>	
curar	<tr>		descendir	<intr>	<de>
cursar	<tr>		descobrir	<tr>	
dansar	<intr>	<0>	descompondre	<vase>	<en>
dar	<tr>		desconèixer	<tr>	
datar	<intr>	<de>	descriure	<tr>	
davallar	<intr>	<0>	descuidar	<tr>	
debatre	<vase>	<amb>	desembocar	<intr>	<en>
decantar	<vase>	<cap>	desencadenar	<tr>	
decidir	<vase>	<a>	desenrotllar	<vase>	<en>
declarar	<tr>		desenvolupar	<vase>	<en>
decorar	<tr>		desfer	<vase>	<de>
decretar	<tr>		desfilat	<tr>	
dedicar	<vase>	<a>	designar	<tr>	
deduir	<vase>	<de>	desitjar	<tr>	
defensar	<tr>		deslligar	<vase>	<de>
definir	<tr>		deslliurar	<vase>	<de>
deformar	<tr>		desmuntar	<tr>	
defugir	<tr>		despenjar	<tr>	
deixar	<tr>		despertar	<tr>	
delimitar	<tr>		desplaçar	<vase>	<a>
demanar	<tr>		desplegar	<tr>	
demostrar	<tr>		desprendre	<vase>	<de>
denominar	<tr>		despullar	<vase>	<de>
denotar	<tr>		destacar	<tr>	
denunciar	<tr>		destinar	<vase>	<a>
depassar	<tr>		destorbar	<tr>	
dependre	<intr>	<de>	destriar	<tr>	
derivar	<intr>	<de>	destrossar	<tr>	
derrotar	<tr>		destruir	<tr>	
desafiar	<tr>		desvetllar	<tr>	

desviar	<tr>		divertir	<vase>	<amb>
desxifrar	<tr>		dividir	<vase>	<en>
detallar	<tr>		divulgar	<tr>	
detectar	<tr>		doblegar	<vase>	<a>
detenir	<tr>		documentar	<vase>	<sobre>
determinar	<tr>		doldre	<tr>	
detestar	<tr>		dominar	<tr>	
deturar	<vase>	<a>	donar	<tr>	
deure	<vase>	<a>	dormir	<intr>	<0>
devorar	<tr>		dotar	<intr>	<0>
dibuixar	<tr>		dreçar	<vase>	<a>
dictar	<tr>		dubtar	<intr>	<de>
diferenciar	<vase>	<de>	dur	<tr>	
diferir	<intr>	<en>	durar	<intr>	<0>
difícultar	<tr>		edificar	<tr>	
difondre	<tr>		editar	<vase>	<a>
diluir	<vase>	<amb>	educar	<tr>	
dinar	<intr>	<0>	efectuar	<tr>	
dipositar	<tr>		eixamplar	<tr>	
dir	<tr>		eixir	<intr>	<de>
dirigir	<vase>	<a>	eixugar	<vase>	<amb>
discernir	<tr>		elaborar	<tr>	
disculpar	<vase>	<per>	elegir	<tr>	
discutir	<tr>		elevant	<vase>	<a>
disfressar	<vase>	<de>	eliminar	<tr>	
disminuir	<tr>		elucidar	<tr>	
disparar	<tr>		embarcar	<intr>	<0>
dispersar	<vase>	<per>	embolcallar	<tr>	
disposar	<intr>	<de>	embolicar	<tr>	
disputar	<vase>	<a>	embrutar	<tr>	
dissenyar	<tr>		emergir	<intr>	<de>
dissimular	<tr>		emetre	<tr>	
dissoldre	<vase>	<en>	emigrar	<intr>	<a>
distingir	<tr>		emmagatzemar	<tr>	
distreure	<tr>		emmalaltir	<intr>	<0>
distribuir	<vase>	<en>	emmarcar	<tr>	

emmudir	<intr>	<0>	enganxar	<vase>	<a>
emocionar	<vase>	<amb>	enganyar	<tr>	
empaitar	<tr>		engegar	<tr>	
emparar	<vase>	<en>	engendrar	<tr>	
empènyer	<tr>		englobar	<tr>	
emplenar	<tr>		engolir	<tr>	
emprar	<tr>		engrandir	<vase>	<per>
emprendre	<tr>		enlairar	<vase>	<a>
empresonar	<tr>		enllaçar	<vase>	<amb>
enamorar	<vase>	<de>	enllestir	<tr>	
encaixar	<intr>	<0>	enlluernar	<tr>	
encalçar	<tr>		enraonar	<intr>	<0>
encaminar	<vase>	<a>	enregistrar	<tr>	
encapçalar	<tr>		enretirar	<tr>	
encarar	<vase>	<amb>	enriquir	<tr>	
encarnar	<tr>		ensenyar	<tr>	
encarregar	<vase>	<de>	ensopegar	<intr>	<amb>
encendre	<tr>		ensorrar	<tr>	
encertar	<tr>		ensumar	<tr>	
encetar	<tr>		entendre	<tr>	
encomanar	<intr>	<0>	enterrar	<tr>	
encomanar	<vase>	<a>	entrar	<intr>	<en>
encoratjar	<tr>		entregar	<tr>	
encreuar	<vase>	<amb>	entrenar	<vase>	<a>
endegar	<tr>		entretenir	<vase>	<a>
enderrocar	<tr>		entreveure	<tr>	
endevinar	<tr>		entrevistar	<vase>	<amb>
endinsar	<vase>	<en>	entusiasmar	<vase>	<amb>
endreçar	<tr>		enumerar	<tr>	
enfadar	<vase>	<amb>	enunciar	<tr>	
enfilat	<vase>	<per>	envair	<tr>	
enfocar	<tr>		envejar	<tr>	
enfonsar	<vase>	<en>	envellir	<intr>	<0>
enfortir	<tr>		enviar	<tr>	
enfosquir	<vase>	<per>	envoltar	<tr>	
enfrontar	<vase>	<amb>	enyorar	<tr>	

equivaler	<intr>	<a>	espiar	<tr>	
equivocar	<vase>	<de>	espolsar	<vase>	<amb>
erigir	<vase>	<en>	esquinçar	<tr>	
errar	<vase>	<de>	esquivar	<tr>	
errar	<vase>	<de>	establir	<tr>	
esbandir	<tr>		estalviar	<tr>	
esborrar	<tr>		estendre	<vase>	<per>
esbossar	<tr>		estimar	<tr>	
esbrinar	<tr>		estimular	<tr>	
escalfar	<vase>	<a>	estirar	<tr>	
escampar	<vase>	<per>	estranyar	<vase>	<de>
escapar	<vase>	<de>	estremir	<vase>	<de>
escaure	<vase>	<en>	estrenar	<tr>	
esclafar	<tr>		estrènyer	<tr>	
esclatar	<intr>	<0>	estructurar	<vase>	<en>
escolar	<vase>	<per>	estudiar	<tr>	
escollir	<tr>		esvair	<vase>	<a>
escoltar	<tr>		esvanir	<vase>	<amb>
escoltar	<tr>		esverar	<vase>	<per>
escombrar	<tr>		evaporar	<vase>	<de>
escometre	<tr>		evidenciar	<tr>	
escopir	<tr>		evitar	<tr>	
escórrer	<vase>	<per>	evocar	<tr>	
escriure	<tr>		evolucionar	<intr>	<0>
escurçar	<tr>		examinar	<tr>	
esgotar	<tr>		excavar	<tr>	
esguardar	<tr>		excedir	<vase>	<en>
esmaltar	<tr>		exceptuar	<tr>	
esmentar	<tr>		exclamar	<vase>	<amb>
esmerçar	<vase>	<en>	excloure	<tr>	
esmorzar	<tr>		executar	<tr>	
esmunyir	<vase>	<per>	exercir	<tr>	
espantar	<tr>		exercitar	<tr>	
espavilar	<vase>	<per>	exhibir	<tr>	
especificar	<tr>		exigir	<tr>	
esperar	<tr>		existir	<tr>	

expandir	<vase>	<per>	fondre	<vase>	<amb>
experimentar	<tr>		foradar	<tr>	
explicar	<tr>		foragitar	<tr>	
explicitar	<tr>		forçar	<tr>	
explorar	<tr>		forjar	<tr>	
explotar	<tr>		formar	<tr>	
exportar	<tr>		formular	<tr>	
exposar	<tr>		fornir	<tr>	
expressar	<tr>		fotografiar	<tr>	
expulsar	<tr>		fotre	<tr>	
exterminar	<tr>		fracassar	<tr>	
extingir	<vase>	<en>	fregar	<tr>	
extreure	<tr>		frenar	<tr>	
fabricar	<tr>		freqüentar	<tr>	
facilitar	<tr>		fruir	<intr>	<de>
fallar	<tr>		fugir	<intr>	<de>
faltar	<tr>		fullejar	<tr>	
fecundar	<tr>		fumar	<tr>	
felicitar	<tr>		funcionar	<intr>	<0>
fer	<tr>		fundar	<tr>	
ferir	<tr>		garantir	<tr>	
festejar	<tr>		gastar	<tr>	
fiar	<intr>	<de>	gaudir	<intr>	<de>
ficar	<vase>	<a>	gemegar	<tr>	
figurar	<intr>	<0>	generalitzar	<vase>	<a>
filar	<tr>		generar	<tr>	
filtrar	<vase>	<per>	gestionar	<tr>	
finalitzar	<tr>		girar	<vase>	<cap>
finançar	<tr>		gosar	<tr>	
fingir	<tr>		governar	<tr>	
firmar	<tr>		gratar	<vase>	<a>
fitar	<tr>		gravar	<tr>	
fixar	<tr>		guaitar	<tr>	
florir	<intr>	<0>	guanyar	<tr>	
fomentar	<tr>		guardar	<tr>	
fonamentar	<vase>	<en>	guarir	<tr>	

guiar	<tr>		iniciar	<tr>	
habitar	<tr>		inscriure	<vase>	<en>
haver	<tr>		inserir	<vase>	<en>
heretar	<tr>		insinuar	<tr>	
honorar	<tr>		insistir	<intr>	<en>
humiliar	<vase>	<per>	inspirar	<vase>	<en>
identificar	<tr>		instal.lar	<vase>	<a>
ignorar	<tr>		instaurar	<tr>	
igualar	<tr>		insultar	<tr>	
il.luminar	<tr>		integrar	<vase>	<en>
il.lustrar	<tr>		intensificar	<vase>	<amb>
imaginar	<tr>		intentar	<intr>	<de>
imitar	<tr>		intercanviar	<tr>	
impedir	<tr>		interessar	<vase>	<per>
implicar	<tr>		interpretar	<tr>	
importar	<tr>		interrogar	<vase>	<sobre>
imposar	<vase>	<a>	interrompre	<tr>	
impregnar	<tr>		intervenir	<intr>	<en>
impressionar	<tr>		introduir	<tr>	
imprimir	<tr>		intuir	<tr>	
improvisar	<tr>		inundar	<tr>	
impulsar	<tr>		inventar	<tr>	
inaugurar	<tr>		invertir	<tr>	
incidir	<intr>	<en>	investigar	<tr>	
incitar	<intr>	<a>	invitar	<intr>	<a>
inclinat	<vase>	<a>	invocar	<tr>	
incloure	<tr>		jeure	<intr>	<0>
incorporar	<vase>	<a>	judicar	<tr>	
incorporar	<vase>	<a>	jugar	<intr>	<0>
incrementar	<tr>		jurar	<tr>	
indicar	<tr>		justificar	<tr>	
induir	<intr>	<a>	jutjar	<tr>	
inflar	<vase>	<de>	lamentar	<vase>	<de>
influir	<intr>	<en>	limitar	<vase>	<a>
informar	<tr>		liquidar	<tr>	
ingressar	<intr>	<a>	llançar	<tr>	

llaurar	<tr>		mesclar	<vase>	<amb>
llegir	<tr>		mesurar	<tr>	
llençar	<tr>		millorar	<tr>	
llepar	<tr>		minvar	<tr>	
llevar	<tr>		mirar	<tr>	
lligar	<tr>		mobilitzar	<tr>	
lliscar	<intr>	<0>	modelar	<tr>	
lliurar	<vase>	<a>	modificar	<tr>	
lloar	<tr>		molestar	<tr>	
llogar	<tr>		morir	<intr>	<0>
lluir	<tr>		mormolar	<tr>	
lluitar	<intr>	<per>	mossegar	<tr>	
localitzar	<vase>	<a>	mostrar	<tr>	
madurar	<tr>		motivar	<tr>	
maldar	<intr>	<per>	moure	<vase>	<de>
maleir	<tr>		mudar	<vase>	<a>
malmetre	<tr>		mullar	<tr>	
manar	<intr>	<0>	multiplicar	<vase>	<per>
mancar	<tr>		muntar	<tr>	
manejar	<tr>		murmurar	<intr>	<0>
manifestar	<vase>	<en>	narrar	<tr>	
manipular	<tr>		navegar	<intr>	<0>
mantenir	<tr>		necessitar	<tr>	
marcar	<tr>		nedar	<intr>	<0>
marxar	<intr>	<0>	negar	<tr>	
mastegar	<tr>		negligir	<tr>	
matar	<tr>		negociar	<tr>	
matissar	<tr>		néixer	<intr>	<0>
meditar	<intr>	<0>	netejar	<tr>	
menar	<intr>	<a>	nodrir	<vase>	<de>
mencionar	<tr>		nomenar	<tr>	
menjar	<tr>		notar	<tr>	
mentir	<intr>	<0>	obeir	<tr>	
menysprear	<tr>		oblidar	<tr>	
meravellar	<vase>	<de>	obligar	<intr>	<a>
merèixer	<tr>		obrar	<intr>	<0>

obrir	<tr>		pastar	<tr>	
observar	<tr>		pasturar	<tr>	
obstar	<tr>		patir	<tr>	
obtenir	<tr>		pecar	<intr>	<de>
ocasionar	<tr>		pegar	<tr>	
ocórrer	<intr>	<0>	penedir	<vase>	<de>
ocultar	<tr>		penetrar	<intr>	<en>
ocupar	<tr>		penjar	<tr>	
odiar	<tr>		pensar	<intr>	<en>
ofegar	<tr>		pentinar	<vase>	<amb>
ofendre	<tr>		percebre	<tr>	
oferir	<tr>		perdonar	<tr>	
oir	<tr>		perdre	<tr>	
omplir	<tr>		perdurar	<intr>	<0>
operar	<intr>	<0>	perfeccionar	<tr>	
opinar	<intr>	<0>	perflar	<vase>	<a>
oposar	<vase>	<a>	perjudicar	<tr>	
oprimir	<tr>		perllongar	<vase>	<fins>
optar	<vase>	<per>	permetre	<tr>	
ordenar	<tr>		perpetuar	<tr>	
organitzar	<tr>		perseguir	<tr>	
orientar	<vase>	<cap>	persistir	<intr>	<0>
originar	<vase>	<en>	pertànyer	<intr>	<a>
oscil.lar	<intr>	<entre>	pertocar	<intr>	<a>
ostentar	<tr>		pertorbar	<tr>	
pagar	<tr>		pescar	<tr>	
palesar	<tr>		pescar	<tr>	
palpar	<tr>		petar	<intr>	<0>
paralitzar	<tr>		picar	<tr>	
parar	<tr>		pintar	<tr>	
parir	<tr>		planar	<intr>	<damunt>
parlar	<intr>	<de>	planificar	<tr>	
participar	<intr>	<en>	plantar	<tr>	
partir	<intr>	<de>	plantejar	<tr>	
passar	<tr>		plànyer	<vase>	<de>
passejar	<vase>	<per>	plaure	<intr>	<0>

plegar	<tr>		presumir	<intr>	<de>
plorar	<intr>	<0>	pretendre	<vase>	<de>
ploure	<intr>	<0>	prevaler	<intr>	<0>
poblar	<tr>		prevenir	<tr>	
poder	<vase>	<de>	preveure	<tr>	
podrir	<vase>	<a>	privar	<intr>	<de>
polir	<tr>		procedir	<intr>	<de>
pondre	<tr>		proclamar	<tr>	
portar	<tr>		procurar	<tr>	
posar	<vase>	<a>	produir	<vase>	<en>
posseir	<tr>		programar	<tr>	
possibilitar	<tr>		progressar	<intr>	<0>
potenciar	<tr>		prohibir	<tr>	
practicar	<tr>		projectar	<vase>	<en>
precedir	<tr>		prolongar	<vase>	<en>
precipitar	<vase>	<a>	prometre	<tr>	
precisar	<tr>		promoure	<tr>	
predicar	<tr>		pronunciar	<tr>	
predir	<tr>		propagar	<vase>	<per>
predominar	<tr>		propiciar	<tr>	
preferir	<tr>		proporcionar	<tr>	
pregar	<tr>		proposar	<vase>	<de>
preguntar	<vase>	<per>	propugnar	<tr>	
prémer	<tr>		prosperar	<intr>	<0>
prendre	<tr>		prosseguir	<tr>	
preocupar	<vase>	<de>	protagonitzar	<tr>	
preparar	<tr>		protegir	<tr>	
prescindir	<intr>	<de>	protegir	<tr>	
presenciar	<tr>		provar	<tr>	
presentar	<tr>		proveir	<vase>	<de>
preservar	<tr>		provenir	<intr>	<de>
presidir	<tr>		provocar	<tr>	
pressentir	<tr>		publicar	<tr>	
pressionar	<intr>	<0>	pujar	<intr>	<0>
pressuposar	<tr>		puntualitzar	<tr>	
prestar	<tr>		punxar	<tr>	

qualificar	<intr>	<de>	refermar	<tr>	
quedar	<vase>	<a>	reflectir	<tr>	
qüestionar	<tr>		reflexionar	<intr>	<sobre>
radicar	<intr>	<en>	reforçar	<tr>	
rajar	<tr>		refredar	<vase>	<amb>
raonar	<vase>	<amb>	refugiar	<vase>	<en>
raure	<intr>	<en>	refusar	<tr>	
reaccionar	<intr>	<amb>	regalar	<tr>	
realitzar	<tr>		regalimar	<tr>	
reaparèixer	<intr>	<0>	regar	<tr>	
rebaixar	<tr>		regir	<tr>	
rebentar	<tr>		regirar	<tr>	
rebre	<tr>		registrar	<tr>	
rebutjar	<tr>		regnar	<tr>	
recaure	<intr>	<en>	regular	<tr>	
recitar	<tr>		reivindicar	<tr>	
reclamar	<tr>		relacionar	<vase>	<amb>
recobrar	<tr>		relliscar	<intr>	<0>
recobrir	<tr>		remarcar	<tr>	
recollir	<tr>		rematar	<tr>	
recolzar	<vase>	<en>	remenar	<tr>	
recomanar	<tr>		remetre	<intr>	<a>
reconèixer	<tr>		remoure	<tr>	
reconstruir	<tr>		remugar	<tr>	
recordar	<tr>		remuntar	<vase>	<a>
recórrer	<tr>		rendir	<tr>	
rectificar	<tr>		renegar	<intr>	<de>
recular	<intr>	<0>	renéixer	<intr>	<a>
recuperar	<tr>		renovar	<tr>	
redactar	<tr>		rentar	<tr>	
redreçar	<tr>		renunciar	<intr>	<0>
reduir	<vase>	<a>	renyar	<tr>	
reeixir	<intr>	<a>	reparar	<tr>	
reemplaçar	<tr>		repartir	<tr>	
refer	<tr>		repassar	<tr>	
referir	<vase>	<a>	repercutir	<intr>	<en>

repetir	<tr>		reunir	<vase>	<a>
replacar	<intr>	<0>	revelar	<tr>	
reposar	<intr>	<0>	revestir	<tr>	
reprendre	<tr>		revifar	<tr>	
representar	<tr>		revisar	<tr>	
reprimir	<tr>		reviure	<tr>	
reproduir	<tr>		riure	<intr>	<0>
requerir	<tr>		robar	<tr>	
resar	<tr>		rodar	<tr>	
reservar	<vase>	<per>	rodejar	<tr>	
residir	<intr>	<en>	rodolar	<intr>	<per>
resignar	<vase>	<a>	romandre	<intr>	<en>
resistir	<tr>		rompre	<tr>	
resoldre	<tr>		rosegar	<tr>	
respectar	<tr>		rumiar	<tr>	
respirar	<tr>		saber	<tr>	
respondre	<intr>	<a>	sacrificar	<tr>	
ressaltar	<tr>		sacsejar	<tr>	
resseguir	<tr>		saltar	<intr>	<0>
ressonar	<intr>	<0>	saludar	<tr>	
ressuscitar	<tr>		salvar	<tr>	
restablir	<tr>		sancionar	<tr>	
restar	<intr>	<0>	satisfer	<tr>	
restaurar	<tr>		satisfer	<tr>	
resultar	<tr>		seduir	<tr>	
resumir	<tr>		segar	<tr>	
retallar	<tr>		seguir	<tr>	
retardar	<tr>		seleccionar	<tr>	
retenir	<tr>		sembrar	<tr>	
retirar	<vase>	<a>	sentir	<tr>	
retornar	<intr>	<a>	senyalar	<tr>	
retratar	<vase>	<a>	separar	<tr>	
retre	<tr>		servar	<tr>	
retreure	<tr>		servir	<tr>	
retrobar	<tr>		seure	<intr>	<0>
retrocedir	<intr>	<0>	signar	<tr>	

significar	<tr>		substituir	<tr>	
simbolitzar	<tr>		succeir	<intr>	<0>
simplificar	<tr>		suggerir	<tr>	
simular	<tr>		sumar	<tr>	
sintetitzar	<tr>		superar	<tr>	
situar	<vase>	<en>	suplir	<tr>	
sobrar	<tr>		suportar	<tr>	
sobrepasar	<tr>		suposar	<tr>	
sobresortir	<intr>	<0>	suprimir	<tr>	
sobreviure	<intr>	<0>	surar	<intr>	<0>
sobtar	<intr>	<0>	suscitar	<tr>	
sofrir	<tr>		suspendre	<tr>	
sol.licitar	<tr>		tallar	<tr>	
soler	<intr>	<0>	tancar	<tr>	
solucionar	<tr>		tapar	<tr>	
somiar	<tr>		tardar	<intr>	<a>
somniar	<tr>		tastar	<tr>	
somriure	<intr>	<0>	teixir	<tr>	
sonar	<tr>		telefonar	<intr>	<a>
sopar	<intr>	<0>	témer	<tr>	
sorgir	<tr>		temptar	<tr>	
sorprendre	<tr>		tendir	<intr>	<a>
sortir	<intr>	<0>	tenir	<tr>	
sospesar	<tr>		tenyir	<vase>	<de>
sospirar	<intr>	<0>	testimoniar	<tr>	
sospitar	<tr>		tibar	<tr>	
sostenir	<tr>		tintar	<intr>	<amb>
sotmetre	<vase>	<a>	tirar	<tr>	
sovintejar	<tr>		titllar	<intr>	<de>
suar	<intr>	<0>	titular	<vase>	<a>
subjectar	<tr>		tocar	<tr>	
submergir	<vase>	<en>	tolerar	<tr>	
subministrar	<tr>		tombar	<tr>	
subratllar	<tr>		topar	<intr>	<amb>
subscriure	<tr>		tornar	<intr>	<a>
subsistir	<intr>	<0>	traçar	<tr>	

tractar	<vase>	<de>	vèncer	<tr>	
traduir	<vase>	<en>	vendre	<tr>	
trair	<tr>		venerar	<tr>	
trametre	<intr>	<0>	venir	<intr>	<a>
tranquil.litzar	<tr>		venjar	<vase>	<de>
transcendir	<tr>		ventar	<tr>	
transcòrrer	<tr>		verificar	<tr>	
transcriure	<tr>		vessar	<tr>	
transferir	<tr>		vestir	<vase>	<de>
transformar	<vase>	<en>	vetllar	<tr>	
transitar	<intr>	<per>	veure	<tr>	
transmetre	<tr>		viatjar	<intr>	<per>
transportar	<tr>		vibrar	<intr>	<0>
traslladar	<vase>	<a>	vigilar	<tr>	
traspasar	<tr>		vincular	<tr>	
travessar	<tr>		violar	<tr>	
treballar	<intr>	<0>	visitar	<tr>	
tremolar	<intr>	<0>	viure	<intr>	<0>
trencar	<tr>		volar	<intr>	<0>
trepitjar	<tr>		voler	<tr>	
treure	<tr>		voltar	<tr>	
triar	<tr>		vorejar	<tr>	
trigar	<intr>	<a>	votar	<tr>	
triomfar	<intr>	<0>	xerrar	<intr>	<0>
trobar	<tr>		xisclar	<intr>	<0>
trontollar	<tr>		xiular	<tr>	
trucar	<intr>	<a>	xocar	<intr>	<amb>
ultrapassar	<tr>		xuclar	<tr>	
unificar	<tr>				
unir	<vase>	<a>			
usar	<tr>				
utilitzar	<tr>				
vacil.lar	<intr>	<0>			
valer	<vase>	<de>			
valorar	<tr>				
variar	<intr>	<0>			

Bibliografia

- Alcina, J. i Blecua, J. M. (1975). *Gramática española*. Barcelona: Ariel.
- Alsina, A., Badia, T., Boleda, G., Bott, S., Àngel Gil, Quixal, M., i Valentín, O. (2002). CATCG: a general purpose parsing tool applied. *A Proceedings of Third International Conference on Language Resources and Evaluation*, Las Palmas, Espanya.
- Bartra, A. (2002). La passiva i les construccions que s'hi relacionen. A Solà, J. (Ed.), *Gramàtica del Català Contemporani*, pàgines 2111–2179. Barcelona: Empúries.
- Bel, A. (2002). Les funcions sintàctiques. A Solà, J. (Ed.), *Gramàtica del Català Contemporani*, pàgines 1075–1147. Barcelona: Empúries.
- Boleda, G. (2003). Adquisició de classes adjectivals. Treball de recerca, Universitat Pompeu Fabra, Barcelona.
- Brent, M. (1991). Automatic acquisition of subcategorization frames from untagged text. *A Proceedings of the 29th Annual Meeting of the ACL*, pàgines 209–214, Berkeley, USA.
- Brent, M. (1993). From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 7, 243–262.
- Bresnan, J. (1982). *The mental representation of grammatical relations*. Cambridge: The MIT Press.
- Bresnan, J. (2001). *Lexical-Functional Syntax*. Oxford: Blackwell Publishers.
- Briscoe, T. i Carroll, J. (1997). Automatic extraction of subcategorization from corpora. *A Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, Washington, USA.
- Cano, R. (1987). *Estructuras sintácticas transitivas en el español actual*. Madrid: Gredos.
- Cano, R. (1999). Los complementos de régimen verbal. A Bosque, I. i Demonte, V. (Eds.), *Gramàtica descriptiva de la lengua española*, pàgines 1221–1279. Madrid: Espasa-Calpe.
- Carroll, J. (1998a). Automatic acquisition of subcategorization frames and selectional preferences from corpora. Talk given at the workshop "Practical Acquisition of Large-Scale Lexical Information" at CSLI, Stanford.
- Carroll, G. i Rooth, M. (1998b). Valence induction with a head-lexicalized PCFG. *A Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, pàgines 36–45, Granada, Espanya.
- Cohen, W. W. (2002). Ripper. <http://www-2.cs.cmu.edu/wcohen/>.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.
- Eagles (1996). Eagles recommendations for the morphosyntactic annotation of corpora. Informe tècnic EAG-TCWG-MAC/R, ILC-CNR, Pisa.
- Eckle, J. i Heid, U. (1996). Extracting raw material for a German subcategorization lexicon from newspaper text. *A Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX'96*, Budapest, Hongria.

- Fabra, P. (1956). *Gramàtica catalana*. Barcelona: Teide.
- Hernanz, M. i Brucart, J. (1987). *La Sintaxis*. Barcelona: Crítica.
- Hindle, D. i Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 19(2), 103–120.
- Huddleston, R. i Pullum, G. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- IEC (1995). *Diccionari de la llengua catalana*. Barcelona: Edicions 3 i 4. Publicacions de l'Abadia de Montserrat.
- Justeson, J.S. i Katz, S. (1995). Principled disambiguation: discriminating adjective senses with modified nouns. *Computational Linguistics*, 17(1), 1–19.
- Karypis, G. (2002). *CLUTO: A Clustering Toolkit*. Manual de CLUTO, versió 2.0.
- Keller, F., Corley, M., Corley, S., Crocker, M. W., i Trewin, S. (1999). Gsearch: A tool for syntactic investigation of unparsed corpora. *A Proceedings of the EACL Workshop on Linguistically Interpreted Corpora*, Bergen, Noruega.
- Lapata, M. (1999). Acquiring lexical generalizations from corpora: A case study for diathesis alternations. *A Proceedings of the 37th Annual Meeting of the ACL*, pàgines 397–404, Maryland, USA.
- Lapata, M. (2000). *The Acquisition and Modeling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy*. Tesi doctoral, University of Edinburgh.
- Lapata, M. i Brew, C. (1999). Using subcategorization to resolve verb class ambiguity. *A Proceedings of WVLC/EMNLP*, pàgines 266 – 274, College Park, USA.
- Levin, B. (1993). *English Verb Classes and Alternations: a preliminary investigation*. Chicago i Londres: University of Chicago Press.
- Levin, B. i Rappaport, M. (1995). *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Cambridge: The MIT Press.
- Lorente, M. (1996). Una proposta de classificació dels verbs catalans. *A Estudis de lingüística oferts a Antoni M. Badia i Margarit*, pàgines 78–101. Barcelona: Departament de Filologia Catalana. Universitat de Barcelona. Publicacions de l'Abadia de Montserrat.
- Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.
- Manning, C. (1993). Automatic acquisition of a large subcategorisation dictionary from corpora. *A Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pàgines 235–242, Columbus, USA.
- Manning, C. D. i Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.
- Matsumoto, Y. (2002). Lexical knowledge acquisition. A Mitkov, R. (Ed.), *Oxford handbook of computational linguistics*. Oxford: Oxford University Press.
- Merlo, P. i Stevenson, S. (2001). Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3), 373–408.

- Morel, J., Vivaldi, J., Cabré, T., Yzaguirre, L. D., i Torner, S. (1997). Etiquetari de l'IULA. Informe tècnic, Institut Universitari de Lingüística Aplicada, Barcelona.
- Pereira, F. C. N., Tishby, N., i Lee, L. (1993). Distributional clustering of english words. A *Proceedings of the 31st Annual Meeting of the ACL*, pàgines 183–190, Columbus, USA.
- Quinlan, J. (1992). *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufman Publishers, Inc.
- Quixal, M. (2003). Theoretical basis and implementation of a linguistic-based morphosyntactic tagger for Catalan. Treball de recerca, Doctorat en Ciència Cognitiva i Llenguatge, Universitat Pompeu Fabra.
- Rafel, J. (1994). Un corpus general de referència de la llengua catalana. *Caplletra*, 17, 219–250.
- Resnik, P. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. Tesi doctoral, Department of Computer and Information Science, University of Pennsylvania.
- Ribas, F. (1995). *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy*. Tesi doctoral, Universitat Politècnica de Catalunya.
- Rooth, M., Riezler, S., Prescher, D., Carroll, G., i Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. A *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pàgines 104–111, Maryland, USA.
- Rosselló, J. (2002). El SV, I: verb i arguments verbals. A Solà, J. (Ed.), *Gramàtica del Català Contemporani*, pàgines 1853–1949. Barcelona: Empúries.
- Sag, I. i Pollard, C. (1987). *Information-based syntax and semantics*. Stanford: Center for the Study of Language and Information.
- Schulte im Walde, S. (1998). Automatic semantic classification of verbs according to their alternation behaviour. Treball de recerca, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart.
- Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation behaviour. A *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pàgines 747–753, Saarbrücken, Alemanya.
- Schulte im Walde, S. i Brew, C. (2002). Inducing German semantic verb classes from purely syntactic subcategorisation information. A *Proceedings of the 40th Annual Meeting of the ACL*, pàgines 223–230, Philadelphia, USA.
- Ushioda, A., Evans, D., Gibson, T., i Weibel, A. (1993). The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. A *The Acquisition of Lexical Knowledge from Text*, pàgines 95–106. Columbus, USA.
- Vallduví, E. (2002). L'oració com a unitat informativa. A Solà, J. (Ed.), *Gramàtica del Català Contemporani*, pàgines 1221–1279. Barcelona: Empúries.
- Vázquez, G. (1997). El clíctic *es* i la diàtesi anticausativa. *Sintagma*, pàgines 61–73.
- Vendler, Z. (1967). Verbs and times. A *Linguistics in Philosophy*, pàgines 97–121. Ithaca i Londres: Cornell University Press.

- Zernik, U. (1989). Lexical acquisition: learning from corpora by capitalising on lexical categories. *A Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pàgines 1556–1564, Detroit, USA.
- Zhao, Y. i Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis. Informe tècnic 01-40, Department of Computer Science & Engineering, University of Minnesota.