UNIVERSITAT DE BARCELONA

Grau de Lingüística

Treball de Fi de Grau

Curs 2018-2019

TÍTOL: Detecció computacional de la metaforicitat adjectival en un nou corpus planià

NOM DE L'ESTUDIANT: Enric Caussa Morales

NOM DEL TUTOR: Irene Castellón Masalles

Barcelona, 14 de juny del 2019

UNIVERSITAT DE BARCELONA

Coordinació d'Estudis
Facultat de Filologia

Gran Via
de les Corts Catalanes, 585
08007 Barcelona

Tel. +34 934 035 594
fil-coord@ub.edu
www.ub.edu

Declaració d'autoria

Amb aquest escrit declaro que sóc l'autor/autora original d'aquest treball i que no he emprat per a la seva elaboració cap altra font, incloses fonts d'Internet i altres mitjans electrònics, a part de les indicades. En el treball he assenyalat com a tals totes les citacions, literals o de contingut, que procedeixen d'altres obres. Tinc coneixement que d'altra manera, i segons el que s'indica a l'article 18, del capítol 5 de les Normes reguladores de l'avaluació i de la qualificació dels aprenentatges de la UB, l'avaluació comporta la qualificació de "Suspens".

Barcelona, a 14 de juny del 2019

Signatura: 

Detecció computacional de la metaforicitat adjectival en un nou corpus planià

Enric Caussa Morales

Curs 2018-2019. Data d'entrega: 14 de juny del 2019

Resum

Les metàfores són un fenomen predominant en la llengua habitual, però també ho són en la llengua literària. La seva detecció pot esdevenir una eina útil per fer l'anàlisi estilomètric d'un autor o una obra i fer-ho de forma computacional permet processar grans volums de dades tot evitant la costosa intervenció humana. A més a més, la seva detecció també pot ser profitosa per una aplicació de comprensió del llenguatge natural. En aquest treball es presenta un classificador que combina tècniques de detecció de metàfores basades en l'estat de la qüestió i es posa a prova en les característiques construccions de nom i adjectiu en un corpus de Josep Pla creat des de zero.

Paraules clau: construcció de corpus, metàfora, processament del llenguatge natural

Metaphors are as prevalent in the daily speech as in any literary context. Detecting them in a text can be a useful measure to carry out a stylometric analysis of a work of literature, and doing so taking advantage of computer algorithms can help reduce the workload of human analysts. Moreover, detecting metaphoricity can be a helpful tool for natural language understanding applications. Here, a competitive classifier based on state-of-the-art techniques will be introduced. It will also be evaluated on the characteristic noun and adjective constructions of a novel Josep Pla textual corpus.

Keywords: corpus construction, metaphor, natural language processing

1 Introducció

En el nostre dia a dia transformem, modelem i adaptem la naturalesa dels conceptes que ens ajuden a descriure el món. En la bibliografia lingüística és habitual llegir que una *llengua* està *amenaçada*, que *competeix* en un *ecosistema lingüístic* i *està en risc d'extinció* quan les condicions són adverses i, que quan es deixa de *transmetre*, la llengua esdevé *morta*. Aquestes combinacions modifiquen el sentit literal de *llengua* i li apliquen qualitats que normalment acompanyen paraules d'altres camps. Exemplifiquen, així, el mecanisme que ens permet traslladar propietats semàntiques d'una paraula a una altra o, en termes més tècnics, d'un domini conceptual a un altre. Aquest mecanisme és la metàfora. Ens permet entendre conceptes mitjançant-ne altres i el trobem de forma tan sistemàtica i recorrent que s'afirma que "pensem metafòricament"(Lakoff & Johnson, 2003).

Si bé la metàfora és un fenomen natural i central en el sistema cognitiu humà, quines metàfores utilitzen els parlants i quins dominis hi tenen lloc varia en funció del bagatge social i cultural de cadascú. Conseqüentment, els trets quantitatius i qualitatius de les metàfores esdevenen interessants objectes d'estudi de qualsevol anàlisi estilístic d'un autor (Bousfield, 2014), essent, doncs empremtes per a la caracterització estilomètrica d'un discurs. Aquesta caracterització havia de ser necessàriament una tasca manual, però actualment es pot fer de en gran part utilitzant mètodes de processament del llenguatge natural, és a dir, programes informàtics que permeten analitzar un *input* lingüístic i transformar-lo en les dades que ens interessen.

Aquest camp és, precisament, on la lingüística computacional pot jugar un paper clau: gràcies als constants avenços informàtics, els investigadors del processament del llenguatge natural (PLN) han pogut implementar eines cada vegada més sofisticades que permeten processar corpus textuals cada vegada més grans, a part de fer servir anàlisis *quantitatives*, les tècniques d'anàlisi *qualitatives* poden ajudar, per exemple, a predir l'edat o gènere d'autors de textos a partir dels contextos en què apareixen les paraules (Bayot & Gonçalves, 2016), descobrir construccions importants durant el període d'adquisició lingüística (Martí, Taulé, Kovatchev & Salamó, 2019) i detectar de forma automàtica la presència i característiques de les metàfores en un escrit (Mykowiecka, Marciniak & Wawer, 2018; Schulder & Hovy, 2014; Shutova, Sun & Korhonen, 2010; Strzalkowski et al., 2013).

Aquesta última tasca és la que es pretén abordar en aquest treball. En primer lloc, es presentarà el rerefons essencial per definir la forma i funció de la metàfora. Després, s'exploraran els diferents camins que s'han seguit per fer una detecció i mesura computacional (DCM) de les metàfores. Finalment, es descriurà un experiment en el qual s'implementaran alguns dels mètodes de detecció de metàfores i es posaran a prova en els parells de nom+adjectiu d'un corpus de les Obres Completes de Josep Pla creat *ad hoc*.

Aquest experiment té per objectiu comprovar dues hipòtesis: (a) els mètodes de detecció de metàfores que combinen regles i estadística són efectius per detectar la metaforicitat adjectival en aquest corpus, i (b) en el corpus de Pla s'hi trobaran un nombre d'adjectius en configuració metafòrica més gran que en ús literal.

2 Definició i aproximacions per a la detecció de metàfores

La detecció computacional de metàfores pot ser una peça clau per poder arribar a la comprensió automàtica del llenguatge natural ("Natural Language Understanding"). Per detectar una metàfora primer cal explicar què s'entén per metàfora, quines característiques poden facilitar-ne la identificació i com s'ha abordat la tasca fins l'actualitat.

2.1 Què és una metàfora?

El fenomen de les metàfores ha estat un objecte d'estudi abordat des de disciplines com ara la filosofia, la ciència cognitiva i, per descomptat, la lingüística. Dins de la lingüística s'ha descrit la metàfora des del punt de vista de la pragmàtica (Bousfield, 2014; Clark, 2014), de la psicolingüística (Glucksberg, 2003) i també des de la semàntica (Fass & Wilks, 1983), entre d'altres. Així, fer un resum o, simplement, llistar totes les aproximacions que s'han fet al respecte és, doncs, una tasca inabarcable (Kirby, 1997). D'aquesta manera, en aquesta secció només es farà una brevíssima aproximació a les definicions que s'han aportat al llarg de la història.

La tradició filosòfica occidental ha percebut la metàfora com a interessant figura retòrica des de fa més de dos mil·lenis. Aristòtil va ser un dels grans filòsofs del llenguatge en l'Antiguitat Clàssica i, entre altres temes relacionats amb la llengua, va tractar la metàfora. A *Poètica* (Aristòtil, 2005: cap. 21), comparava la metàfora amb un trasllat d'una paraula a una altra. Aquest moviment té lloc dins d'un mateix nivell taxonòmic o bé en pot implicar un altre. Aquests nivells taxonòmics són el *genus* i l'*espècie*, fent un paral·lelisme amb el món de la biologia i reconeixent, doncs, qualitats ontològiques de les metàfores. A més a més, descrivia d'una altra forma de metàfora: l'analogia. Aquesta consisteix en una relació que implica relacions entre quatre entitats. Per exemple, *el vespre de la vida* és una analogia en què "l'edat avançada és a la vida el que el vespre és al dia". Per Aristòtil, la metàfora era un "do" (Kirby, 1997: p. 534) que permet percebre la semblança, anomenar allò que no té denominació, aporta vivacitat, i que té qualitats de dolçor, claredat i foraneitat (ib. p. 541). El filòsof grec li donava una importància especial a l'impacte retòric de la metàfora. Observava, també que la metàfora era un important instrument de persuasió, d'aprenentatge i de transmissió d'idees.

Una altra destacada visió de la metàfora és la de Friedrich Nietzsche, qui qüestionava el rol tradicional de la metàfora de mecanisme retòric i afirmava que els processos metafòrics són, ben al contrari, centrals en el significat (Hinman, 1982: p. 182). Segons el filòsof, la metàfora té el seu començament en el procés de percepció: el sol fet de veure quelcom ho transforma en un estímul nerviós, que a la vegada passa a ser una imatge mental, fet que és *per se* metafòric. Quan ho expressem a "una imatge, imitada per un so", afirmava Nietzsche, tornem a passar d'una representació ("esferes") a una altra (ib. p. 184). De forma similar, quan fem una comparació, o bé englobem conceptes dins de categories, estem igualant allò que és desigual, i allò que implica els sentits requereix un "trasllat metonímic" a una causa externa imaginària. Argumenta, també, que aquelles metàfores que ocorren freqüentment són candidates a fossilitzar-se en la llengua comuna, i que acaben perdent el seu caràcter metafòric: són el que s'anomena "metàfores mortes"("dead metaphors"). El filòsof alemany compara aquest fet amb "les monedes que, amb l'ús, perden les seves imatges, i ara no són més que metall; ja no són monedes" (ib. p. 189).

Una de les perspectives contemporànies més citades és a *Metaphors We Live by* de Lakoff i Johnson (2003). En aquest tractat, els autors van refutar la idea de la metàfora com un do, tal com defensava Aristòtil. Ben al contrari, emfasitzaven el caràcter *normal* i habitual de la metàfora i la van situar en el centre de la maquinària lingüístico-cognitiva, tot mostrant evidència psicolingüística (vg. ib. capítol *afterword 2003*).

Caracteritzaven el fenomen de la metàfora principalment des de la seva vessant conceptual, com a mecanisme que permet entendre un concepte a través d'un altre, tot mitjançant unes relacions sistemàtiques que formen una ontologia. Aquesta ontologia, de forma similar als *genus* i espècies d'Aristòtil, forma una jerarquia que estructura els sistemes que permeten crear metàfores de forma productiva. Aquests sistemes són les *metàfores conceptuals* i corresponen a l'estructura semàntica comuna a les metàfores lingüístiques.

Com a exemple de metàfora conceptual, Lakoff i Johnson van proposar, entre moltes altres, EL TEMPS ÉS UN RECURS VALUÓS. Aquesta formaria part de la jerarquia conceptual com a node que englobaria EL TEMPS ÉS DINERS. Uns exemples de metàfores lingüístiques que formarien part d'aquesta jerarquia serien (adaptats d'ib. p. 10):

(1) Estic *perdent* el temps

(2) Amb aquest dispositiu podràs *estalviar* temps

(3) He *invertit* massa temps en ella

En el model de de Lakoff i Johnson, els mecanismes que permeten la jerarquia de les metàfores conceptuals es basen en la similitud i en la co-ocurrència entre dominis conceptuals i es manifesten en propietats interaccionals de l'experiència humana. Així, en l'exemple de metàfora conceptual que hem vist es fa un símil entre DINERS i TEMPS.

Com a metàfora conceptual de co-ocurrència els autors proposaven MÉS ÉS AMUNT (ib. p. 156).

Hem vist com la metàfora és essencial per a la comprensió, però també ho és per a la producció. Lakoff i Johnson reiteren que els parlants creen noves metàfores sense que se n'adonin, contradient, doncs, l'èmfasi aristotèlic en la vessant retòrica de la metàfora. Creant noves metàfores, asseguren, es creen noves connexions en les dimensions naturals de l'experiència que, a la seva vegada, creen noves formes d'organització i comprensió de l'experiència. En una metàfora lingüítica, però, sempre hi haurien unes qualitats que amb la metàfora quedarien afegides o ressaltades i algunes que es perdrien o quedarien emmascarades.

Aquestes connexions poden ser característiques d'individus i de cultures. Per exemple, els autors de la influent obra afirmen que, en la cultura occidental, els *arguments* i les *opinions* es *defensen* i s'*ataquen*, lligant, així, els trets conceptuals de l'ARGUMENTACIÓ amb els de la LLUITA. Plantejaven que aquestes connexions no són inherents en aquests conceptes sinó que són creades per la nostra cultura i que les societats actuen segons la seva forma de veure el món. De forma inversa, doncs, convidaven a imaginar una societat que, amb una percepció de l'argumentació diferent, actuaria de forma que no tindria res a veure amb l'occidental.

Hem vist com la correlació permet *conceptualitzar* entitats a través d'altres gràcies a la relació de dominis conceptuals, però també permet la *referència*. En aquest últim cas, però, quan s'utilitza una entitat per ocupar el lloc d'una altra (Lakoff i Johnson, 2003, p. 36) no parlem de metàfora sinó de metonímia. Segons Lakoff i Johnson, la metonímia té, principalment, una funció referencial i, també, una funció de comprensió. El seu ús, així com el de la metàfora, és la norma i no l'excepció, és inconscient i, està profundament lligat amb la llengua del dia a dia. Per exemple, és habitual substituir els continguts pels continents, com a una pregunta que pot ocòrrer en una trobada social qualsevol: *em pots passar l'aigua?*. Hi han, també, altres referències possibles, com ara PRODUCTE PER MARCA, CONTROLADOR PER CONTROLAT, etc (ib. p. 39) però la tendència general de funcionament metonímic és la contigüitat (Steen, 2007, p. 70). La diferent funció de metàfora i metonímia no implica que quan hi hagi una no pot haver-hi l'altra; en una mateixa frase es pot donar una co-ocurrència de metonímia i metàfora. Un exemple d'aquest tipus de frase seria "ja veig el que vols dir", on hi ha una metàfora conceptual ENTENDRE ÉS VEURE i una metonímia de TOT PER LA PART.

El model de Lakoff i Johnson, si bé és generalment acceptat per la bibliografia consultada, té dures crítiques, especialment pel que fa a la dificultat d'establir un sentit literal (Wierzbicka, 1986). A més, també hi ha la problemàtica de l'ontologia de les metàfores estructurals, així com la incapacitat de les metàfores per definir conceptes abstractes (ib. p. 291).

A (Steen, 2007: cap. 3), es presenten models de la metàfora alternatius al model

de les metàfores conceptuals (CMM, d'ara endavant), amb el qual comparteixen trets. Per exemple, el model dels múltiples espais de Fauconnier i Turner (ib. p. 52) funciona, de manera semblant al CMM amb una projecció d'un domini *source* a un *target*. La diferència principal, però, és que la projecció ocorre no en dos dominis, sinó en dos espais conceptuals, a més d'un tercer espai genèric i un quart espai de "mescla" ("*blended space*"). Aquests quatre espais, en comptes de funcionar com a contenidors conceptuals com el model bi-domini, serien estructures que contenen "conceptualitzacions de l'experiència més aviat específiques" (ib. p. 51). D'aquesta manera, cada espai aportaria qualitats semàntiques al parell metafòric, cosa que facilitaria la comprensió de metàfores amb relacions complexes.

Un altre model per explicar la metàfora és el d'*inclusió en classe* de Glucksberg (ib. p. 64). En aquest model les metàfores funcionen com a mecanismes d'atribució de trets semàntics. Aquesta atribució es dóna en el moment que una paraula passa d'una categoria conceptual *source* a una categoria conceptual *target* a través d'una categoria superordinada que té els trets semàntics que se li volen atribuir. Aquestes categories superordinades són creades *ad hoc* i amb l'ús esdevenen convencionals. El fet de la creació sota demanda de noves categories superordinades contrasta amb el model de Lakoff i Johnson, el qual fonamentava la maquinària de la metàfora en una base neural i, per tant, inherent en la capacitat lingüística. En canvi, segons el model d'inclusió en classe, no és requereix la pre-existència de metàfores conceptuals en la ment dels parlants: les noves metàfores, així com les categories superordinades es creen de forma productiva, i la comprensió de les metàfores es faria mitjançant, principalment, la desambiguació semàntica.

Hi ha, però, una teoria de la metàfora que construeix un pont entre aquest últim model, el de la inclusió en classe, i el CMM de Lakoff i Johnson. Aquest model, de (Gentner Steen, 2007, p. 54), defensava que, efectivament, en una construcció metafòrica es fa una comparació de dominis. Això, però, només passaria en les noves metàfores; en les metàfores convencionalitzades el mecanisme de funcionament seria el de d'inclusió en classe. Així, segons aquest model, una nova metàfora *a és x* es produiria necessàriament gràcies a una comparació entre dos dominis. Aquesta comparació crearia, com a resultat, una categoria metafòrica. A mida que la metàfora vagi creixent en ús i es combini amb nous elements *target* (*a és x*, *b és x*, etc.) evocant la mateixa significació metafòrica, la relació acabaria convencionalitzada. En aquest cas, l'element font x esdevindria polisèmic, amb el seu significat *intra*-domini original, així com el seu significat *inter*-domini adquirit. Una futura comprensió o producció d'un ús metafòric de x implicaria únicament el mecanisme d'inclusió en classe (Bowdle & Gentner, 2005).

Aquests diferents models de metaforicitat impliquen diferents factors que s'hauran de cercar durant la tasca de detecció de metàfores. Per exemple, si el marc de treball inclou el model de metàfores conceptuals de Lakoff i Johnson, es podrà detectar una en el cas

que es detecti l'activació de dos dominis conceptuals així com una relació de *mapping* entre ells (ib. p. 75). Si s'opta pel model d'inclusió en classe, el mètode de detecció variarà en funció de quin tipus de metàfora sigui l'objecte d'identificació. Així, per les noves metàfores, s'haurà de dur a terme una detecció de les dues categories implicades: la de destí i la base, la qual activa la categoria superordinada que recull les dues. En tot cas, sigui quin sigui el model que s'esculli, en una tasca de detecció computacional de metàfores s'haurà de treballar sobre unitats processables per un algoritme computacional.

2.2 Treballs anteriors

Per tal de veure de quina manera ha estat operacionalitzada la tasca de detecció computacional de metàfores (DCM, d'ara endavant), en aquesta secció s'exploraran diferents enfocaments que es poden dur a terme en aquesta tasca.

Durant els anys 80 i 90 del segle passat la tendència general per a la DCM era la de desenvolupar bases de dades de coneixement lingüístic que s'aprofitarien durant l'execució de regles de detecció metafòrica. A (Fass, 1991) es fa una vista general dels treballs que s'havien dut a terme i descriuen, entre altres, el model de selecció, d'interacció, comparació i el de les metàfores conceptuals. Abreujadament, el model de comparació utilitzava com a heurística la discriminació entre una metàfora i un sentit literal. La comparació la planteja com una operació que funcionava a través de l'abstracció, l'analogia, el símil i l'anomalia (ib. p. 51), i es valora mitjançant un algoritme que analitza els trets semàntics dels termes implicats en la metàfora. Segons el model d'interacció, una metàfora es pot detectar com una operació de *mapping* de vocabulari d'un domini a un altre que implica certa coherència lògica entre la metàfora en sí i un conjunt de frases del domini *source*.

Aquest model d'interacció és al qual s'associa Fass i segons el qual opera el seu mètode de DCM, el met*. met* planteja la DCM com un problema de classificació en quatre categories: significat literal, metafòric, metàfora i metonímia i utilitzen un algoritme basant-se en relacions de restricció semàntica. Segons Fass, la restricció semàntica és aquell atribut que posseixen els sentits d'"algunes classes de *parts of speech*". Per exemple, en una metàfora s'estaria trencant la preferència entre els dominis *source* i *target*. D'aquesta manera, l'autor crea una sèrie de *frames* a mà. Utilitzant les relacions amb un algoritme que cerca restriccions semàntiques anòmales entre dos *frames* i depenent del tipus de relació de restricció que es trenca o es respecta, es distingeix entre un sentit literal, una metàfora i una metonímia (ib, p. 84). Ara bé, pel fet de cercar relacions semàntiques anòmales per poder arribar a detectar la metàfora, l'algoritme sobregeneralitza i considera metafòriques i anòmales relacions que en principi només són anòmales com ara *the car drank coffee* (ib., p. 76).

Schulder i Hovy (2014) proposen un camí alternatiu per a la tasca de DCM motivats per la dificultat d'obtenir dades exhaustives dels dominis conceptuals que es volen analit-

zar. Així doncs, desenvolupen un mètode per trobar la magnitud de com una peça lèxica es pot considerar "fora de lloc", basant-se en l'assumpció de les restriccions semàntiques (tal com argumentava Fass, 1991). El mètode consisteix en una mesura estadística, TF-IDF, que compara la freqüència relativa d'una paraula tenint en compte el domini *target* que es mesura i el conjunt de dominis en general. Aquesta mesura assigna valors baixos tant a paraules massa freqüents en els dominis en general com a les paraules de molt baixa freqüència en els dominis en general i, de forma inversa, valors alts a elements que ocorren en pocs dominis. Els autors adapten la TF-IDF per filtrar aquelles paraules que ocorren amb massa freqüència, quedant-se, doncs, amb aquelles paraules que es poden considerar metafòriques o, més concretament, fora de lloc respecte un cert domini. La informació sobre quines paraules corresponen a quin domini l'extreuen a partir d'un corpus que recopila mil milions de pàgines del WWW, ClueWeb09. En fan una mostra i la subdivideixen en 8 dominis temàtics, que utilitzaran per derivar la TF-IDF d'alguns mots clau i desenvolupen dos classificadors que etiqueten com a sí/no-metafòric aquelles paraules fora de domini amb valors TF-IDF per sobre del llindar. El primer classificador utilitza únicament la mesura TF-IDF i obté un recall mitjà del 0.53 i una precisió mitjana del 0.25. El segon classificador utilitza, a més de la TF-IDF, paràmetres com etiquetes PoS, i obté un resultats inversos: 0.65 de precisió i 0.26 de recall en el millor dels casos.

Tal com hem vist, la creació *ad hoc* de bases de dades pot ser útil per a la DCM. A partir dels anys 90 van començar a sorgir projectes on grups de recerca codificaven el coneixement lingüístic en forma de repositoris digitals que es poden utilitzar de forma lliure per a la recerca. WordNet (Miller, Beckwith, Fellbaum, Gross & Miller, 1990), un exitós exemple de projecte d'aquest tipus, consisteix en una xarxa de *synsets*. Aquests *synsets* recullen conjunts de sentits lèxics i s'enllacen en forma de xarxa amb altres *synsets* tot recollint relacions d'hiperonímia i hiponímia, sinonímia i antonímia, meronímia, etc.

La xarxa semàntica WordNet és un dels recursos que s'utilitza a Krishnakumaran i Zhu, 2007, on desenvolupen un algoritme que anota com a si/no metafòrics parells de nom+adjectiu, de verb+nom i metàfores copulatives amb verb ser i nom. Utilitzen com a base de coneixement les cadenes d'hiperònims de WordNet i apliquen l'algoritme, sobre el corpus Berkeley Master Metaphor List. Aconsegueixen una precisió general del 70% i *recall* del 61%.

FrameNet i PropBank són dos exemples més de projectes que recullen dades semàntiques. Basat en la teoria de *frames* de Fillmore, FrameNet (Baker, Fillmore & Lowe, 1998) proporciona un lexicó de patrons de realització d'events representat en els marcs semàntics. Aquests, a la seva vegada, recullen aquells dominis que es poden realitzar en una determinada frase. A més, els dominis tenen una organització jeràrquica on es relacionen amb *frames* subordinats o superordinats. Per exemple, en el frame de TRANSPORT s'inclou el de CONDUIR i aquest inclou abstraccions com ara CONDUCTOR, PASSATGER, VEHICLE, etc.

PropBank (Kingsbury & Palmer, 2003), és un corpus que, per la seva banda, recull verbs amb sentits desambiguats segons la seva estructura argumental i la relació entre els arguments, tot oferint certa compatibilitat amb els *synsets* de WordNet. Ofereix, també, verbs desambiguats segons els rols semàntics, així com arbres amb la informació semàntica i sintàctica codificada en anàlisis de dependències.

D'aquestes bases de coneixement, se'n poden extreure bases de dades del seu coneixement expert, de la manera que millor s'adeqüi a les necessitats dels experiments. Així, a (Gedigian, Bryant, Narayanan & Ciric, 2006) utilitzen, en primer lloc, un algoritme basat en regles que relaciona els *marcs* de FrameNet a sentits de PropBank i, en segon lloc, un classificador estadístic per distingir entre sentits verbals metafòrics i literals en un subconjunt del corpus del Wall Street Journal. La diferència entre els sentits literals i els metafòrics requeia en l'estructura argumental del conjunt metafòric. D'aquesta manera, comparaven les estructures argumentals del verb. Si aquesta estructura argumental no coincidia amb la del FrameNet, llavors aquell verb tenia possibilitats de trobar-se en una configuració metafòrica. Descobreixen que, en aquest subconjunt, prop del 90% dels verbs s'estan utilitzant de forma metafòrica i aconsegueixen una precisió del 0.92.

La combinació de mètodes basats en regles i mètodes estadístics, com a Gegidian i altres, és el mètode de DCM en un corpus de poesia pel qual s'opta a (Kesarwani, Inkpen, Szpakowicz & Tanasescu, 2017). Les regles que utilitzen Kesarwani i coll·laboradors són la concreció-abstracció, extreta a partir de la cadena d'hiperònims de WordNet i termes relacionables a partir de relacions de ConceptNet, una base de dades que agrega informació de WordNet, Wikipedia, Wiktionary. Utilitzen, també, mesures estadístiques com ara la PMI i la similitud cosinus. Totes les mesures estadístiques i els valors de les regles s'incorporen a un model que acompanya els elements als quals volen detectar metàfores. Aquest model és la base de l'entrenament d'un classificador estadístic, que anota la metaforicitat de construccions de tipus còpules *x és y*, parells de verb i nom i nom i adjectiu extretes del corpus de poesia. Obtenen 0.797 de precisió i 0.860 de *recall*.

Els mètodes estadístics són la base pel treball de Shutova i col·laboradors (2010), on utilitzen les tècniques de *clustering* per detectar la metaforicitat a través del comportament estructural de verbs i noms. La tasca de *clustering* agrupa aquelles paraules que comparteixen similitud semàntica, com és el cas de conceptes concrets (com ara aliments), i agrupa per associcacions amb un mateix domini d'origen en el cas de conceptes abstractes. D'aquesta manera, parteixen d'un corpus de sintagmes verbals i utilitzen una tècnica de clustering espectral per trobar les preferències de selecció d'aquests verbs així com per *clusteritzar* els noms. Comparant la relació entre verb i nom, comproven si s'ha produït una violació de les restriccions de selecció, de forma similar al que hem vist en Fass, 1991. Utilitzant aquest mètode, sobrepassen el 79% de precisió en la tasca de DCM.

La tasca de *clustering* resulta ser una eina potent per la DCM, tal com hem vist a Shutova et al 2010 i com es va comprovar més tard a (Strzalkowski et al., 2013), on

utilitzen, a més, una nova mesura: la imaginabilitat nominal. En aquest treball combinen el *clustering* amb la quantificació de la imaginabilitat nominal, extreta, d'una banda del corpus MRCPD, i, de l'altra, de relacions de WordNet. La imaginabilitat és aquella magnitud que determina la facilitat i rapidesa en què una paraula evoca una imatge mental (ib., p. 70). Segons els autors, els candidats amb més imaginabilitat són els que amb més probabilitat seran metàfores. Així doncs, en aquest treball Strzalkowski i col·laboradors construeixen un diagrama de dependències per un text i construeixen la cadena tema-rema. Eliminen tots aquells elements que formen part d'aquesta cadena assumint que aquests no poden ser candidats a metàfora. Posteriorment, seleccionen les paraules amb més imaginabilitat i, si aquesta sobrepassa un cert llindar, etiqueten el candidat com a metafòric. Comparant el rendiment de l'algoritme amb un *golden standard* d'anotadors humans, obtenen un 0.74 de precisió.

Dunn (2014), tot basant-se en estudis de la lingüística cognitiva contemporània, afirma que la metaforicitat és un tret que es pot medir amb un escalar, és a dir, que no és una distinció binària si-metafòric o no-metafòric. Per implementar un sistema de mesura de la metaforicitat a nivell de frase, extreu un conjunt de frases d'un corpus subdivit en categories temàtiques i desenvolupa un corpus de frases anotats com a no, poc, bastant, o molt metafòrics. A partir de la variabilitat de les anotacions, assigna un valor escalar de 0 a 1 a cadascuna de les frases. Després, elabora un llistat de propietats com ara animacitat, concreció-abstracció i funcionalitat dels objectes (segons els autors, un tornavís, per exemple, és una entitat especialment funcional, mentre que una pedra no ho és). Utilitzant tècniques estadístiques, se'ls assignen valors de metaforicitat a cadascuna d'áquestes propietats i es validen amb els valors escalars del corpus anteriorment anotat. Finalment, l'autor compara un algoritme de DCM binari amb un algoritme DCM escalar amb un subcorpus de validació basat en el VU Amsterdam Metaphor Corpus, obtenint un valor-F comparable en els dos algoritmes: 0.608 en l'éscalar i 0.638 en el binari.

A partir de l'any 2013, com veurem a continuació, apareixen cada cop més treballs que aprofiten models vectorials com ara Word2Vec, de Mikolov i col·laboradors (Mikolov, Chen, Corrado & Dean, 2013) o GloVe Pennington, Socher i Manning, 2014 per analitzar els contextos de les paraules mitjançant els models de semàntica distribucional. Així, són eines que s'utilitzen com ara a (Agres, McGregor, Rataj, Purver & Wiggins, 2016). En aquest article, els autors proposen un model d'evaluació de metàfores utilitzant els models d'espais vectorials que ha de discriminar un ús metafòric o literal en díads verb+nom. Aquests díads provenen d'un corpus que està anotat amb magnituds de semanticitat, familiaritat, metaforicitat i probabilitat d'aparició (Cloze). Utilitzant dos models vectorials: un del Word2Vec i un d'un del Conceptual Discovery Model (CDM), aconsegueixen correlacionar la similitud cosinus (CS) del model Word2Vec amb les magnituds abans esmentades. Així, com menor era la CS, més tendència hi havia a que un díad fos metafòric, especialment en el cas de noves metàfores. El seu model CDM era més flexible per

la tasca de metaforicitat per poder trobar sols els aspectes més metafòrics dels vectors desitjats.

(Do Dinh & Gurevych, 2016) desenvolupen un sistema de xarxes neurals de tipus perceptró multicapa tot aprofitant el Corpus de Metàfores de la VU Amsterdam i el *dataset* Word2Vec que va aportar (Mikolov et al., 2013). El seu mètode de DCM consisteix, en primer lloc en adaptar el *dataset* Word2Vec re-entrenant-lo amb etiquetes de concreció, metaforicitat i de PoS. Ja que aquest corpus està subdividit en gèneres, com ara converses, notícies, prosa o ficció, fan un entrenament independent per a cadascun d'ells, tasca que es fa en una tercera part dels subcorpus, repartint la quarta part restant en un subcorpus de desenvolupament i un de validació. Comprovant el rendiment de la seva xarxa neural, observen que, en el millor dels casos, prop d'un terç de les metàfores del gènere de notícies s'etiqueten correctament. En el gènere que més baix valor-F recull és en el de ficció (0.52, precisió 0.57 i *recall* 0.47).

Bizzoni i col·laboradors (2017) investiguen la metaforicitat en parelles de nom i adjectiu. Parteixen de la hipòtesi que els models de semàntica distribucional són capaços de codificar més informació que la similitud semàntica. Així, utilitzen com a base un model Word2Vec entrenat sobre un corpus de més de 4 mil milions de *tokens* i el re-entrenen amb l'anotació de metaforicitat d'un corpus de més de 8000 parells de nom+adjectiu. Aquestes parelles corresponen a 23 adjectius diferents distribuits en 8 camps semàntics. El reentrenament el fan amb tres arquitectures que operen de manera diferent sobre els vectors del model i comproven l'eficàcia dels seus models sobre un corpus de validació que consisteix en un subconjunt del corpus de metàfores. Obtenen més del 0.86 de precisió i 0.77 de recall en totes les proves que realitzen, cosa que els porta a investigar els trets que poden relacionar-se amb l'èxit del model. Creuen que un dels factors pot ser que el model "aprengui" el nivell d'abstracció dels adjectius, cosa que es pot relacionar amb les investigacions psicolingüístiques que citen.

En resum, hem vist diferents mètodes de DCM que, si bé generalment aconsegueixen detectar si una construcció és metafòrica, les tècniques heurístiques subjacents mostren molta variabilitat. Per exemple, els anàlisis de DCM primerencs, com ara el de Fass, es basaven en regles. A partir de mitjans dels 90, i amb l'augment de capacitat computacional dels ordinadors, les tècniques es decantaven, cada vegada més, per aprofitar els algoritmes estadístics.

3 Experiment de detecció computacional de metàfores

En aquesta secció descriurem un experiment en què s'han posat a prova alguns dels mètodes de detecció de metàfores anteriorment presentats. Consisteix en l'etiquetació de

metaforicitat en parells de nom i adjectiu d'un corpus de les Obres Completes de Josep Pla, creat específicament per aquest experiment.

El projecte d'anàlisis computacional de l'adjectivació en l'obra de Pla, així com el procediment de processament amb eines de vectors foren oferts per la Professora Antònia Martí del Departament de Filologia Catalana i Lingüística General de la UB.

Aquest experiment té com a hipòtesis: (a) els mètodes de detecció de metàfores que combinen regles i estadística són efectius per detectar la metaforicitat adjectival en aquest corpus, i (b) en el corpus planià, s'hi trobaran un nombre d'adjectius en configuració metafòrica més gran que en ús literal.

La primera hipòtesi està motivada per la oportunitat d'implementar diferents mesures que, combinades, siguin capaces de predir la metaforicitat de parells nom+adjectiu. Per implementar aquestes mesures farà falta posar a prova tècniques i coneixements adquirits, especialment, durant els cursos de semàntica, lingüística de corpus i lingüística computacional.

La segona hipòtesi es fonamenta en l'anàlisi de l'adjectivació planiana defensada per Xavier Pla (1997). Concretament, Xavier Pla situa l'estil literari de l'escriptor empordanès en la *ficció autobiogràfica*, que té el "realisme sintètic" (ib. p. 212 i següents) com una de les tècniques predilectes de descripció. Aquesta tècnica consisteix en un ús minimista d'adjectius que defuig de l'anàlisi exhaustiu de la realitat. Així, "el mirall que l'escriptor pretendria passejar al llarg del camí no reflecteix pas la realitat" (ib., p. 267), sinó que opta per la subjectivació per oferir la seva visió personal del món que l'envolta i de les seves experiències. A més a més, l'adjectivació en el realisme sintètic cerca fer una síntesi dels trets tot donant un caràcter subjectiu a la descripció amb un ús eminent d'adjectius relacionats amb la percepció. Podria ser, doncs, que la metàfora adjectival és predominant en l'obra de Josep Pla.

3.1 Metodologia

En aquest apartat es descriuen les tècniques i mètodes que s'han utilitzat per la construcció del corpus de Josep Pla i el general del català, del *golden standard* i de l'algoritme de DCM.

3.1.1 Construcció del corpus de Pla

Per a l'execució d'aquest experiment ha estat necessària, naturalment, l'obtenció d'un corpus de textos produits per Josep Pla. Fins aquest moment no s'havia fet cap anàlisi computacional de l'obra del prolífic escriptor, així que el corpus s'ha hagut de compilar des de zero.

El punt de partida van ser els 46 PDF[1] corresponents a cadascun dels volums de

[1]Facilitats molt amablement per la Càtedra Josep Pla de la Universitat de Girona i Grup62.

l'Obra Completa de Josep Pla. Aquests arxius PDF contenen, d'una banda, la imatge de cadascuna de les pàgines i, d'altra banda, un text obtingut amb el reconeixement òptic de caràcters (OCR) de les imatges. Un programa d'OCR es basa en la transformació d'una imatge en una cadena de caràcters i la qualitat del seu *output* dependrà, de forma proporcional, de la qualitat de la imatge. Malauradament, el text incrustat en els PDFs manifestava defectes d'escaneig, cosa que va motivar la recreació l'OCR.

Per aconseguir, doncs, una versió del text pla a partir de les imatges dels PDF, va ser necessària l'extracció de les imatges que conformaven els arxius. Aquestes imatges, un cop manipulades, es van enviar a Tesseract versió 4, un programa d'OCR, per crear una nova versió del text. Aquesta tasca es va realitzar amb el fitxer d'entrenament proporcionat pel projecte[2]. El resultat va ser un millor reconeixement però encara s'hi manifestaven errors.

Les noves cadenes de text, organitzades per volum, es van concatenar per construir un únic arxiu per a cada volum. Abans, però, feia falta eliminar les metadades de cada pàgina, com ara número de pàgina, títols de seccions, índexs, i tota la informació que, en definitiva, no es pot considerar text original planià. Aquestes tasques es van poder realitzar amb les eines de UNIX `head`, `tail`, `cat` i `grep` i, en alguns casos, va ser necessària intervenció manual.

Un cop obtingut un text que representés cada arxiu, va fer falta una nova fase de processament per eliminar els errors més regulars. `perl` va ser l'eina adequada per aquesta tasca i la seva execució es va dur a terme mitjançant regles de substitució basades en expressions regulars. Així, per a cada línia el programa `perl` cercava una cadena de text que, en cas d'èxit, es substituïa per una altra cadena de text determinada. Per exemple, en la sortida de l'OCR les eles geminades es van reconèixer erròniament com a *H*. Per tant, l'ordre per corregir-ho correspon a {s/([aeiou])H([aeiou])/\1l·l\2/s}. Altres tasques per les quals es va recòrrer a les expressions regulars van ser interrogants detectats com a número dos, signes d'exclamació detectats com a ela o te, cometes angulars detectades com ce, be, be baixa, etc.

També va ser necessària la concatenació de les línies, tasca per la qual s'havia de tenir en compte l'eliminació de guionets de final de línia. Aquesta tasca ja no es va poder dur a terme amb expressions regulars, ja que la guionització pot separar paraules que porten guió, com ara *baliga-balaga*, *donar-te* o *tres-cents*, i també pot marcar el final de línia, cas en el que s'ha d'eliminar el guionet per recuperar el text original. Així, per eliminar o preservar els guionets de final de línia es va utilitzar la llibreria `hunspell` pel llenguatge de programació Python. Aquesta llibreria de correcció ortogràfica conté, també, un potent analitzador morfològic, i es va utilitzar per determinar si es podia eliminar el guionet de final de línia: si la paraula sense guió era normativa (per exemple, *voluminós* és correcte

[2]Tesseract, així com el fitxer d'entrenament utilitzat, `trained_best/ca.tessdata`, es pot descarregar lliurement a https://github.com/tesseract-ocr/

en català normatiu, mentre que *donarme* no ho és), llavors el guió es podia eliminar.

Un cop concatenades les línies, es va utilitzar el programa FreeLing (Padró & Stanilovsky, 2012) per tal d'aconseguir els *tokens* de les cadenes de text. Es va optar pel *tokenitzador* de FreeLing per estar basat en regles de segmentació robustes. Els *tokens* es van passar per l'analitzador ortogràfic `hunspell` per determinar si eren paraules normatives. Degut a la gran quantitat de noms propis i topònims que no apareixen en el lexicó del programa, va ser necessària la creació d'una base de dades, tasca per la qual es va recòrrer a un llistat d'unigrames del corpus de la *Viquipèdia* que es descriurà més tard. Aquest llistat d'unigrames consisteix en una *tokenització* del corpus de la Viquipèdia. Els unigrames es van ordenar per freqüència per tal que es trobessin més ràpidament.

La correcció ortogràfica va ser el següent pas en la construcció d'aquest corpus. En aquesta etapa es trobaven els candidats per substituir una *token* amb forma no normativa. Per trobar-los, es va utilitzar novament `hunspell`, que rebia una *token* i retornava una llista de candidats per a la correcció. Aquesta llista no proporciona candidats amb ordre decreixent segons la probabilitat de ser la forma çorrecta", així que es va desenvolupar un algoritme similar a (Schierle, Schulz & Ackermann, 2008) per trobar el candidat més probable. En primer lloc, l'algoritme utilitza la *Positive Pointwise Mutual Information* (Manning i Schütze, 1999, p. 178), una mesura estadística $\text{PPMI}(x;y) \equiv max(0, \log_2 \frac{p(x,y)}{p(x)p(y)})$ que dóna un valor *ppmi* en funció de la probabilitat de co-ocurrència del terme x seguit de y. Aquest valor PPMI informa de la mesura en què dos termes co-ocorren en relació a la ocurrència independent dels dos termes.

Les probabilitats de co-ocurrència es van calcular a partir del corpus de la Viquipèdia, cosa que va implicar que alguns termes que sortissin en el corpus de Pla no apareguessin en el corpus de la Viquipèdia i viceversa, a més que les probabilitats fossin esbiaixades a favor del corpus de l'enciclopèdia.

En segon lloc, es van ordenar els candidats segons la seva PMI. Després es va calcular la distància de Levenshtein, que indica el número passos que s'ha de seguir per canviar una cadena de text a una altra. Els candidats es van filtrar si la seva distància de Levenshtein respecte el pressumpte error tipogràfic era menor a 2.

Aquest procediment va ajudar a corregir errors d'OCR com ara `iníatigablement` a *infatigablement* o d'`interpel'lació` a *interpel·lació*. No obstant, hi va haver casos on no s'obtenien resultats adequats, ja sigui per baixa xifra d'ocurrències de certs termes al corpus de la Viquipèdia, el biaix inherent obtingut per l'elecció d'aquest corpus, manca de resultats del corrector ortogràfic o simplement un algoritme que no tractava adequadament segons quins casos. Per exemple, aquest algoritme no va ser apte per corregir, per exemple, `SaTuna` a *Sa Tuna*, `vvhishy` a *whisky*, de `P er` a *Per*, etc.

Un cop acabada la tasca de correcció ortogràfica, es va fer una nova passada del conjunt de mòduls de FreeLing. Aquesta vegada, però, es van utilitzar més característiques que en etapes anteriors:

- El *tokenitzador* va convertir el text pla en llistes de paraules.
- Amb l'*splitter* es van convertir les llistes de paraules en llistes de frases que, a la seva vegada, contenien les llistes de paraules.
- L'analitzador morfològic carrega les llistes de paraules i utilitza submòduls –generalment basats en regles com expressions regulars– i específics per cada tasca diferent. Per exemple, es va utilitzar un submòdul per afegir etiquetes PoS a la puntuació, un altre per a la detecció de xifres i dates, un altre per al reconeixement de *Named Entities*, etc. Gràcies a aquest mòdul es van poder normalitzar les dates i xifres a @, agregar col·locacions en forma d'unitats multiparaula (com ara *pa de pessic*, *a causa de*, *en cas de*, etc.).
- Anotació de sentits, que retorna una llista de *synsets* de WordNet. Aquesta llista es passa per l'algoritme de desambiguació semàntica UKB (Padró, Reese, Agirre & Soroa, 2010) per escollir el resultat més probable. Aquest algoritme consisteix en escollir el sentit més probable utilitzant el mètode estadístic PageRank (ib. p. 7)
- L'etiquetatge de Part-of-Speech (PoS), en què es va triar l'algoritme estadístic per defecte, `hmm`.

L'aplicació de FreeLing va donar com a resultat llistes de frases, on cada llista contenia, a la seva vegada, vectors de *token*, lema, identificador de *synset* i etiqueta PoS (basada en el tagset EAGLES). Aquesta informació es va combinar de manera que les frases fossin llistes amb aquest format: `lema/synset/PoS`.

Tot seguit, es va aplicar el models de semàntica vectorial Word2Vec mitjançant la llibreria GenSim (Rehurek & Sojka, 2011). El Word2Vec (Mikolov et al., 2013), entrenat amb 100 dimensions, finestra de paraules 5, .001 de *sampling*, els valors per defecte. Word2Vec és un algoritme estadístic no supervisat que es basa en la *hipòtesi distribucional*. Aquesta hipòtesi prediu que les paraules en contextos similars tenen tendència a mostrar significats similars (Jurafsky i Martin, 2014, capítol 6). Els models de semàntica vectorial posen en pràctica aquesta hipòtesi, i permeten calcular, entre altres mesures, la similitud semàntica.

La similitud semàntica permet saber en quina mesura podem trobar dues o més paraules en un context semblant, i d'aquí, determinar fins a quin punt es poden relacionar. Per exemple, si es busquen les 10 paraules més semblants a *tramuntana* en el model Word2Vec del corpus de Josep Pla s'obté una llista que conté els noms dels vents *mestral*, *garbí*, *migjorn*, *gregal*, *xaloc*, però que també inclou *vent*, així com *nord-oest* i *mistral*, una forma dialectal de *mestral*. Si es col·loquen aquestes paraules en un mateix context, resulten frases perfectament possibles:

(4) Avui bufa una forta *tramuntana*

(5) Avui bufa un fort *migjorn*

(6) ? Avui bufa un fort *nord-oest*

(7) Avui bufa un fort *mestral*

(8) Avui bufa un fort *gregal*

(9) Avui bufa un fort *xaloc*

(10) Avui bufa un fort *garbí*

(11) Avui bufa un fort *vent*

Aquest llistat de frases demostra dues coses: la primera és l'eficàcia del model Word2Vec per obtenir paraules similars: aquestes frases són molt similars. La segona és que, encara que similaritat semàntica i sinonímia semblin termes equivalents, de fet no ho són. Una de les propietats de la sinonímia és que permet intercanviar sinònims en una frase sense canviar el valor de veritat d'una frase. En canvi, aquesta llista presenta un hiperònim *vent* amb cohipònims *tramuntana*, *migjorn*, etc. Els cohipònims comparteixen valor semàntic amb el seu hiperònim i aquest valor amb els seus cohipònims.

3.1.2 Construcció del corpus de la Viquipèdia

Per a la correcció del text resultant de l'OCR va fer falta aconseguir una gran quantitat de text, de manera que es recollissin col·locacions variades, noms propis i topònims que poguessin sortir en el text de Josep Pla, etc. Per aquesta raó, es va escollir el bolcat de març del 2019 de la Viquipèdia[3]. Aquests bolcats consisteixen en arxius XML amb un format propi que contenen tant el text dels articles com les seves metadades.

Per construir una llista d'n-grames és necessària la utilització de text pla. Convertir directament formats que no siguin text pla mostraria soroll en els unigrames. Per tant, va ser necessària la conversió del format propi de la Viquipèdia al text pla.

En una primera instància es va optar, novament, per eliminar les dades extra de l'XML amb expressions regulars. No obstant, el llenguatge de marcat Wiki conté, també formatament, taules, categories, plantilles, etc., cosa que va motivar l'ús del programa especialitzat WikiExtractor. Aquest utilitza regles de substitució basades en expressions regulars, així com màquines d'estat finit per analitzar el marcatge que presenta més complexitat, com ara els enllaços.

El resultat del processament va ser un conjunt de fitxers que es van passar per `grep` per tal d'eliminar línies amb metadades. La concatenació d'aquest arxiu va donar, segons `wc`, un conjunt de 871.927.312 paraules.

D'aquest corpus es va obtenir un llistat d'unigrames passant els arxius concatenats a través del tokenitzador de FreeLing. Utilitzant `tail` es va poder eliminar la primera línia

[3]Disponibles lliurement a https://dumps.wikimedia.org/cawiki/.

del llistat d'unigrames, creant un nou llistat amb un *offset* d'una línia. Amb `paste` per unir els dos arxius i `tr` per convertir la tabulació en un espai, es va poder transformar aquesta unió en un llistat de bigrames. Els unigrames van servir per afegir verdaders positius durant la tasca de verificació ortogràfica del corpus de Pla i, els bigrames, per calcular la PMI per ordenar els candidats per a la tasca de correcció.

Tot seguit, FreeLing va servir per recrear la llista d'operacions, excepte anotació semàntica, que es van dur a terme en el corpus de Pla, és a dir, la *tokenització*, *splitting*, anàlisi morfològica, lematització i *tagging*. Així, es va obtenir un llistat de frases, en què cadascuna era a la seva vegada una llista amb un format semblant al de Pla: `lema/PoS`.

Finalment, es va construir un model Word2Vec amb els mateixos paràmetres que el del corpus de Josep Pla, és a dir, 100 dimensions, sampling de .001 i finestra de 5 elements. Aquest es va enriquir amb una part del corpus CTILC.

3.1.3 Construcció del corpus *golden standard*

Per a al desenvolupament de l'algoritme de detecció de metàfores, una tasca de classificació binària, va ser necessària la cerca d'aquells trets que poguessin ser decisius per a aquesta tasca. Així doncs, es va aconseguir una mostra de 250 parelles nom+adjectiu del corpus de Pla extretes amb expressions regulars que retornaven les parelles més una finestra de 10 paraules.

Aquesta mostra va ser introduïda en una fulla de càlcul i va ser anotada per dos anotadors amb coneixements de lingüística. El protocol que es va seguir per l'anotació va ser el següent:

> Considerem que un nom té una sèrie de trets i que l'adjectiu que l'acompanya pot modificar-los, però també en pot afegir de nous. Si creus que l'adjectiu està afegint nous trets semàntics al nom tenint en compte el context, anotarem la parella com a metafòrica. En cas contrari, l'anotarem com a no metafòrica.

Els casos, és a dir, cada parella nom+adjectiu i el seu context, ocupaven les files. A les columnes, una per a cada anotador, s'hi va introduir el valor segons la percepció de metaforicitat. Després d'una primera anotació, es va realitzar una sessió d'intervenció per comentar dubtes i refinar les instruccions de la tasca, per exemple, es va tenir en compte la diacronia de l'adjectiu. Es van introduir els nous valors de metaforicitat segons els nous criteris. Va ser durant la fase d'anotació on es van detectar alguns errors de *PoS tagging*, on es va detectar la forma d'infinitiu del verb ésser com a nom. Tot seguit es mostra una taula abreujada, que conté les parelles, el context (truncat) i els valors post-sessió d'intervenció.

Parella nom+adjectiu	A1	A2	Context
embat/N furiós/A	s	s	el embat furiós de el materialisme
esforç/N apologètic/A	s	s	esforç apologètic de prat tendir a demostrar
mateix/A efecte/N	n	n	consistir el país fer el mateix efecte
home/N madur/A	n	s	veure un home madur i un noia jove ser
permanent/A sublimitat/N	s	s	en un estat de permanent sublimitat
partit/N català/A	n	n	el contacte de tot el partit català popular
fraseologia/N cientificista/A	n	n	de el món amb el fraseologia cientificista...
ocell/N negre/A	n	n	aquest ocell negre petit net eixut
mandra/N profund/A	s	s	respondre a un mena de mandra profund...

Finalment, es va calcular l'acord entre anotadors comparant les els casos en què els anotadors coincidien en la seva resposta i els que no. A (Steen, 2007, p. 125) es proposa la mesura *kappa* de Cohen, que mesura l'acord entre anotadors cas a cas, és a dir, per a cas parella nom+adjectiu, i té en compte, per una banda, l'*agreement* que es pot donar aleatòriament així com l'acord que s'observa a la mostra. La primera mesura *kappa* de Cohen que es va obtenir va ser κ=.635 (amb 76% d'acord), que va pujar substancialment després de la sessió d'intervenció: κ=.710 (80%), mesura que implica un notable acord. Aquest corpus, amb un 46% de metàfores acordades pels dos anotadors, va esdevenir el *golden standard*.

3.1.4 Desenvolupament de l'algoritme de detecció de metàfores

El *golden standard* es va dividir en dues parts: el corpus de validació i el de desenvolupament. El corpus de desenvolupament consistia en 110 parelles nom+adjectiu i, el corpus de validació, 137 parelles.

L'algoritme que es va desenvolupar per aquest experiment tracta la detecció de metaforicitat adjectival com a problema de classificació binària. És a dir, rep un *input* de parells de metàfores nom+adjectiu i retorna com a *output* el valor boleà VERITAT en cas que la parella es consideri metafòrica i FALS en cas contrari.

El seu mètode de classificació s'inspira en l'utilitzat a (Kesarwani et al., 2017) i (Shutova et al., 2010). Funciona amb quatre mòduls, que recullen una mesura diferent:

1. Positive-Pointwise Mutual Information (PPMI) i la *add-2 Laplace smoothing PPMI* (PPMI "suavitzada"), que consisteix en afegir una constant k als comptadors d'occurència abans de calcular la PPMI (Jurafsky i Martin, 2014, p. 118). Així, amb aquesta combinació es comparar la mesura en què co-ocurren les parelles nom+adjectiu (NA) en els dos corpus tenint en compte, també, les que no hi apareixen. També es va aprofitar una propietat de la PPMI: dóna valors esbiaixats en situacions de manca de dades, cosa que no es dóna en el cas de la PPMI suavitzada. La tendència era que, en els parells metafòrics, la PPMI era més gran en el corpus

de Pla que en el de la Viquipèdia, mentre que la PPMI suavitzada era o bé similar o bé menor en el de Pla pel fet de corregir un baix número de contextos.

2. Les mesures vectorials es van obtenir amb diferents funcions que proporciona la llibreria de semàntica vectorial GenSim. Concretament, es van utilitzar la similitud cosinus comuna, que mesura la diferència entre dos vectors (ib., p. 112) i la *similitud cosinus relativa*, que mesura la similitud relativa entre un vector $\vec{x}$ els n vectors més similars a $\vec{y}$. La tendència esperada era una menor similitud en parells metafòrics que en parells no metafòrics. Això, però, no sempre es respectava, de manera que la similitud es va aprofitar només per refinar algunes regles.

3. A mida que s'analitzava cada parella NA, utilitza el corpus de la Viquipèdia pre construir clusters dels adjectius que coocorren amb el nom. Així, els vectors del model Word2Vec es converteixen en n *clusters* semàntics utilitzant l'algoritme d'aprenentatge automàtic no supervisat *K-means* de la llibreria Scikit-learn (Pedregosa et al., 2011). Així, un cop obté els vectors, K-means busca la forma de trobar un número establert de punts intermedis dins de cada grup de vectors. En aquesta regla es va utilitzar 8, el valor per defecte, com el número de clústers, amb la idea de trobar una "anomalia" de selecció adjectival per part del nom del corpus de Pla i de la Viquipèdia.

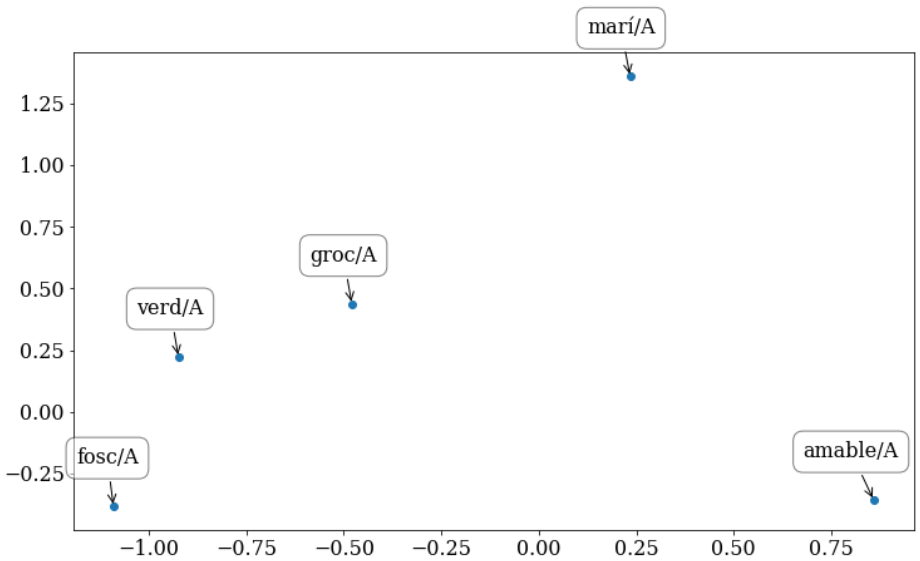

Figura 1 – Un diagrama de dispersió que representa el clústering dels adjectius més freqüents co-ocorrents amb *blau* al corpus de la Viquipèdia. A la part inferior dreta, *amable*, que forma part d'un parell metafòric al corpus planià, està allunyat dels altres nodes.

4. La regla concreció-abstracció es va construir a partir de la xarxa semàntica de WordNet, mitjançant les relacions paradigmàtiques que es poden explorar amb NLTK. Així, es cercava la presència del node "entitat física"o "entitat abstracte"a la cadena d'hiperònims del *synset* que havia etiquetat FreeLing. Segons la bibliografia consultada, la tendència era que les metàfores tenien l'element *target* concret i el *source* abstracte. Aquesta regla, així com la de similitud, rarament s'utilitzava per determinar la metaforicitat; simplement era una puntuació "extra".

Per trobar els llindars de cadascuna d'aquestes regles es van extreure 10 parelles del corpus de desenvolupament segons la seva anotació de metaforicitat, de manera que s'obtinguessin 5 metàfores i 5 no-metàfores. El criteri de metaforicitat, per provenir del *golden standard*, es va considerar la variable independent. Les variables dependents eren PPMI, concreció-abstracció, les similituds cosinus i la clusterització. D'aquesta manera, es van poder trobar les fronteres que dividien les metàfores de les no-metàfores.

Una vegada obtinguts els llindars de cada regla, s'havia de cercar la importància de cadascuna d'elles. Aquesta importància es va codificar en un "pes" numèric que es suma o es resta a una puntuació. Si aquesta puntuació supera un llindar, provoca que l'algoritme consideri que la parella és metafòrica. Com que cada regla contribueix de forma diferent a la classificació de metaforicitat, per obtenir un *recall* o una precisió concretes, es va cercar els valors òptims dels llindars amb les 100 parelles del corpus de desenvolupament restant.

3.1.5 Avaluació

L'avaluació del classificador es va dur a terme amb el càlcul de les mesures estadístiques de precisió i *recall* entre el corpus de validació i l'*output* del classificador d'acord amb Manning i Schütze (1999). Aquestes mesures, provenen del camp de l'Extracció d'Informació i es poden aprofitar per evaluar l'adequació dels models de PLN (ib., p. 267). Mesuren la relació entre dos conjunts: el *seleccionat* i el *target*. D'una banda, el conjunt target conté les avaluacions presents en un *golden standard* i que, idealment, hauria d'obtenir un sistema de classificació. D'altra banda, el conjunt seleccionat conté el conjunt de classificacions dutes a terme pel sistema de classificació. De la intersecció, subtracció i disjunció d'aquests conjunts s'obté el que es coneix com a falsos positius, falsos negatius, verdaders positius i verdaders negatius:

Aquells casos en què el *golden standard* i el sistema de classificació coincideixen són els casos *verdaders* i aquells en què no coincideixen són els *falsos*. Així, els *verdaders positius* són els casos etiquetats com a positius en el *golden standard* i classificats com a positius pel sistema de classificació. Els *verdaders negatius* són els casos en què tant el *golden standard* com el sistema de classificació coincideixen en considerar el cas com a negatiu.

$$\text{Precisió} = \frac{V_p}{V_p + F_p} \qquad \text{Recall} = \frac{V_p}{V_p + F_n} \qquad \text{valor-F} = 2\frac{P \times R}{P + R}$$

Figura 2 –
Les fòrmules de precisió, *recall* i valor-F, on Vp i Vn són verdaders positius i negatius respectivament. Fp i Fn significa falsos positius i falsos negatius respectivament.

Quant als casos *falsos*, els *falsos positius* són classificats com a positius, però que en el *golden standard* són negatius. Si en el *golden standard* fossin positius i, en canvi, negatius segons el sistema de classificació, es tractaria de *falsos negatius*.

Segons (ib., p. 268), la precisió es calcula dividint els verdaders positius per la suma de verdaders positius i falsos positius. Expressa la proporció en què un classificador "encerta" la resposta. El recall, la proporció de positius que es classifiquen com a tal, es calcula amb la divisió dels verdaders positius i la suma dels verdaders positius i falsos negatius.

La precisió i el recall són mesures que estan en tensió: millorar la precisió implica generalment que baixi el recall, i viceversa. En una tasca com la de l'experiment, basada en regles i llindars, doncs, augmentar els llindars per obtenir millor precisió implicarà sacrificar recall. Aquestes dues mesures també es poden combinar en un sola mesura: el valor-F. Es calcula amb la mitjana harmònica de la precisió i el recall i s'ajusta amb un biaix que afavorirà a la precisió o al *recall* en la mesura que s'acosti a 0 o a 1. El valor més habitual –i utilitzat en aquest experiment–, és α=0.50 (ib. p. 269), que dóna igual pes a una i altra mesura.

Així, es va realitzar una passada del classificador sobre els elements que conformaven el corpus de validació (de 137 parelles de nom+adjectiu) i es va avaluar la coincidència entre els resultats del classificador i els del *golden standard*.

3.2 Resultats i discussió

Com en els treballs anteriorment descrits, s'han calculat la precisió (pr), *recall* (re) i valor-F (f) a partir de l'evaluació de dos classificadors de prova. El primer classificador etiqueta la metaforicitat aleatòriament i obté {pr=.38, re=.46, f=.41}. El segon classificador marca tots els casos com a metafòrics. Obté aquestes mesures: {pr=.43, re=1.00, f=.60}. El classificador desenvolupat per aquest treball millora substancialment aquestes mesures amb {pr=.71, re=.57, f=.63} optimitzant per obtenir la màxima precisió i {pr=.48, re=.88, f=.62} optimitzant afavorint el recall. Aquests dos resultats confirmen la primera hipòtesi de l'éxperiment: que els mètodes que s'han comprovat són útils per a la tasca de DCM en un corpus de Josep Pla.

Segons el classificador, en el context de l'obra planiana, hi ha un total de 119.731 parelles nom+adjectiu, de les quals un 59% són metafòriques, Es confirma, doncs, la segona

hipòtesi de l'experiment, que defensa que es trobarien més construccions nom+adjectiu metafòriques que no metafòriques. Aquest resultat valida amb un anàlisi basat en dades la tesi de la corrent literària del realisme sintètic, on es situa Pla, que buscava dir més amb menys tot passant la realitat pel prisma de la subjectivitat.

Quant als *pesos* que es van aplicar a les mesures, el primer resultat, el de màxima precisió, va donar més pes en el comportament en *cluster* que en les mesures basades en la PPMI, i mesures baixes pels dos mòduls restants. El màxim recall es va aconseguir donant el màxim pes a la PPMI, pes mitjà al comportament en cluster, seguit de biaixos mínims per a les mesures vectorials i a les de concreció-abstracció. Si bé la PPMI va resultar ser una mesura útil per detectar possibles candidats de metàfora *novella*, no permetia la discriminació entre un parell metafòric i un que no ho fos.

L'algoritme detecta reeixidament aquelles metàfores novelles com ara *embat furiós*, *llibertat antiga* o *finesa llànguida* gràcies a una nul·la co-ocurrència en el corpus de la Viquipèdia, clusters radicalment diferents i una alta diferència de distàncies vectorials entre les similituds cosinus del corpus de Pla i de la Viquipèdia. Com a falsos positius, l'algoritme es basa en les anomalies que "observa"a *paella valenciana*, *fraseologia cientifista* o *detonació llunyana*. En aquests casos concrets, el mòdul basat en *clustering* detectava una diferència significativa entre els clusters "preferits"per *paella*, *fraseologia* o *detonació* i els que es donaven en els contextos planians.

Així doncs, es demostra la utilitat de la PPMI com un tipus de "filtre"per trobar candidats a noves metàfores, així com la idoneïtat de combinar-la amb anàlisis més qualitatius com la vectorització i clusterització, que permeten observar el çomportament"de noms i adjectius. Amb aquestes tres tècniques solsament, es pot realitzar un anàlisi que no necessita cap dada feta a mà com ara les bases de dades expertes WordNet o FrameNet. En aquest treball, però, no va ser possible utilitzar únicament una mesura: durant el desenvolupament es va obervar que la PPMI, donava falsos positius i que el mòdul de clustering i vectors ajudaven a discriminar la metaforicitat dels parells segons el *golden standard*.

Si bé en aquest treball s'han utilitzat eines, com ara WordNet que parteixen del coneixement lingüístic d'experts, en un futur treball es podria arribar a un coneixement exclusivament basat en mètodes empírics. Aquests podrien observar les ocurrències lingüístiques *in vivo* i determinar el coneixement lingüístic de manera *bottom-up*. L'ús de mesures estadístiques no limiten de cap manera el model de la metàfora sobre el qual es treballa, més aviat el contrari: es poden tractar aquestes mesures abans esmentades com a mètode per trobar "anomalies"o trencaments de restriccions semàntiques, preferències semàntiques o, també, per arribar a una aproximació de xarxes de metàfores conceptuals del model de Lakoff i Johnson. En aquest últim cas, per exemple, es podria fer una clusterització dels adjectius que s'utilitzen metafòricament per trobar els camps semàntics en els quals s'engloben i, d'aquí, ara ja si, potser faria falta intervenció humana per derivar

les metàfores conceptuals a partir dels camps semàntics o paraules afectades..

La manca de segons quins contextos adjectivals ha sigut un factor que no ha jugat un paper decisiu en els resultats de l'algoritme: tot i ser un corpus de prop de 9 milions de *tokens*, dins dels quals s'inclouen composicions com *aromàtico-primaveral*, *literàrio-musical*, falsos positius, com el verb *ésser* en infinitiu detectat com a nom, o errors tipogràfics, bona part dels adjectius ocorren amb baixa freqüència. Concretament, els elements etiquetats com a adjectius i amb major número d'ocurrències són *seu*, amb prop de 59000 ocurrències i *gran*, amb 26000. La mitjana d'ocurrències, però és 41, manifestant, doncs, una presència gran d'adjectius de baixa freqüència.

Si bé en principi una manca d'ocurrències provoca una vectorització de baixa qualitat, que degrada el rendiment del model vectorial, durant el desenvolupament i posterior evaluació de l'algoritme no va semblar que això afectés massa els resultats. Per exemple, dins del corpus de desenvolupament hi havia construccions amb adjectius de baixa freqüència com ara *galopant* o *pendular* en les parelles *taquicàrdia galopant* o *llei pendular*. Els dos adjectius ocorren 18 i 14 vegades respectivament, cosa que en principi afecta negativament la qualitat dels vectors que els representen al model vectorial. Com a conseqüència d'una vectorització esbiaixada, la inclusió d'un adjectiu amb vectors de baixa qualitat afectarà, doncs, el cluster en què s'incorpora i, per tant, la precisió del classificador. En els casos d'aquestes construccions, però, es va poder classificar la metaforicitat, gràcies a contextos regulars, i, especialment, amb la PPMI i les regles basades en la concreció-abstracció de WordNet.

Hi ha hagut casos en què l'algoritme classifica metonímies com a metàfores, com ara en *establiment fosc* (on és *fosca* la llum de *dins* de l'establiment en cas que s'interpreti *establiment* com a edifici). Alternativament, hauria sigut interessant descartar metonímies abans de classificar una parella de nom+adjectiu com a metafòrica. Com a exemples de classificació de metonímia com a metàfora s'ha donat el cas d'*establiment fosc* o *home ferreny*, entre altres. En casos com ara *ocell negre* o *americana blanca*, el mòdul basat en clusterització classifica correctament el parell pel fet de disposar de les seleccions adjectivals preferents d'aquests substantius. També es classifiquen com a metàfores que es poden considerar simultàniament metàfora i metonímia com *pell judaica*.

Quant a les regles basades en les dades de WordNet, els seus resultats no han estat decisius pel rendiment acceptable del classificador. Més aviat, introduïen tant falsos negatius com falsos positius, ja fos per un tractament erroni de la xarxa, com per manca de dades en els nodes propers al *synset*, com per un mal funcionament de la regla que determinava la metaforicitat. Per exemple, en el mòdul de regles de WordNet, es va donar el valor de metaforicitat a una construcció que en el *golden standard* no era metafòrica: *taquígraf autèntic*. En aquest cas concret, el programa no va poder trobar un camí entre el *synset* de *taquígraf* i *autèntic*. Ara bé, per considerar la primera com a entitat física i la segona com atribut abstracte, va donar un valor positiu de metaforicitat d'acord amb

les tendències observades durant el desenvolupament del classificador.

Encara que l'algoritme funciona correctament, hi ha moltes vies que es poden seguir per millorar els seus valors. Per començar, com a arquitectura, en comptes de cercar els pesos manualment, en una implementació alternativa (i més similar a l'estat de la qüestió actual), s'aprofitaria la intel·ligència artificial per construir un classificador més precís. Així, una xarxa neural o algun altre tipus d'arquitectura d'aprenentatge automàtic podria utilitzar les dades i classificacions que rebria. Per realitzar aquesta tasca i entrenar adequadament el model, però, un corpus de metàfores anotat per experts hauria de contenir més de 250 parelles. Així, seria convenient disposar d'un corpus de metàfores anotat com el de la VU Amsterdam però pel català. L'algoritme en sí també es podria millorar aprofitant més models a part del Word2Vec com ara el GloVe, i complementar-los amb altres mesures estadístiques per fer una anàlisi lingüístic bottom-up. Aquesta anàlisi es podria dur a terme no només a partir de les parelles nom+adjectiu, sinó que també podria ampliar-se a totes aquelles construccions que poden ser metàfores en potència.

4 Conclusió

Per acabar, doncs, s'han examinat els factors que poden ajudar a detectar computacionalment les metàfores. Aquests factors són, en especial, la PPMI i les tècniques basades en vectors. També s'ha aplicat reeixidament l'algoritme, basat en regles i estadística en un corpus de Josep Pla que, si bé presentava deficiències al principi, es va poder corregir, com a mínim en part. De fet, aquest és, fins on es té constància, el primer treball que analitza computacionalment la metàfora en l'obra completa de Josep Pla.

Com a propostes per a treballs futurs, es podrien fer millores en l'arquitectura del classificador per assemblar-se més a l'éstat de la qüestió. També es podria deixar de fer ús de bases de dades expertes per passar a un ús exclusivament estadístic.

El següent pas a seguir seria verificar la utilitat del classificador amb parells que no siguin nom+adjectiu, com ara construccions copulatives o bé verbals. Finalment, es podria comprovar-lo amb un corpus general del català, tasca per la qual seria de profit tenir un corpus de metàfores del català.

5 Agraïments

En primer lloc, vull donar les gràcies a la Professora Antònia Martí, no només per la idea sinó també per la forma d'abordar la tasca d'analitzar l'ádjectivació en l'obra de Pla. A la Professora Irene Castellón, per tutoritzar-me molt pacientment, per donar-me interessants idees i consells, ajudar-me a anotar el corpus, i orientar-me en el laberint de la bibliografia. Sense elles no podria haver executat aquest treball. En segon lloc,

agraeixo al Professor Xavier Pla de la Càtedra Josep Pla de la UdG el seu entusiasme amb aquest projecte, la cessió dels textos de Pla i el regal dels llibres interessantíssims.

Referències

Agres, K. R., McGregor, S., Rataj, K., Purver, M. & Wiggins, G. A. (2016). Modeling metaphor perception with distributional semantics vector space models. A *C3GI@ ESSLLI*.

Aristòtil. (2005). *Aristotle: Poetics* (J. Sachs, Trad.). Estats Units: Focus Publishing R. Pullins Co.

Baker, C. F., Fillmore, C. J. & Lowe, J. B. (1998). The Berkeley Framenet Project. A *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 86 - 90). Association for Computational Linguistics.

Bayot, R. K. & Gonçalves, T. (2016). Author Profiling using SVMs and Word Embedding Averages. A *CLEF (Working Notes)* (pp. 815 - 823).

Bizzoni, Y., Chatzikyriakidis, S. & Ghanimifard, M. (2017). Deep Learning: Detecting Metaphoricity in Adjective-Noun Pairs. A *Proceedings of the Workshop on Stylistic Variation* (pp. 43 - 52).

Bousfield, D. (2014). Stylistics, speech acts and im/politeness theory. A M. Burke (Ed.), *The Routledge Handbook of Stylistics* (1a ed., pp. 118 - 135). Routledge Handbooks in English Language Studies. Regne Unit: Routledge.

Bowdle, B. F. & Gentner, D. (2005). The career of metaphor. *Psychological review*, *112*(1), 193.

Clark, B. (2014). Stylistics and relevance theory. A M. Burke (Ed.), *The Routledge Handbook of Stylistics* (1a ed., pp. 155 - 174). Routledge Handbooks in English Language Studies. Regne Unit: Routledge.

Do Dinh, E.-L. & Gurevych, I. (2016). Token-level metaphor detection using neural networks. A *Proceedings of the Fourth Workshop on Metaphor in NLP* (pp. 28 - 33).

Dunn, J. (2014). Measuring metaphoricity. A *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 745 - 751).

Fass, D. (1991). met*: a method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, *17*(1), 49 - 90.

Fass, D. & Wilks, Y. (1983). Preference Semantics, Ill-formedness, and Metaphor. *Comput. Linguist. 9*(3-4), 178 - 187. Recuperat des de http://dl.acm.org/citation.cfm?id=1334.980082

Gedigian, M., Bryant, J., Narayanan, S. & Ciric, B. (2006). Catching metaphors. A *Proceedings of the Third Workshop on Scalable Natural Language Understanding* (pp. 41 - 48). Association for Computational Linguistics.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in cognitive sciences*, *7*(2), 92 - 96.

Hinman, L. M. (1982). Nietzsche, metaphor, and truth. *Philosophy and phenomenological research*, *43*(2), 179 - 199.

Jurafsky, D. & Martin, J. H. (2014). *Speech and language processing*. Pearson London.

Kesarwani, V., Inkpen, D., Szpakowicz, S. & Tanasescu, C. (2017). Metaphor detection in a poetry corpus. A *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 1 - 9).

Kingsbury, P. & Palmer, M. (2003). Propbank: the next level of treebank. A *Proceedings of Treebanks and lexical Theories* (Vol. 3). Citeseer.

Kirby, J. T. (1997). Aristotle on metaphor. *American Journal of Philology*, *118*(4), 517 - 554.

Krishnakumaran, S. & Zhu, X. (2007). Hunting Elusive Metaphors Using Lexical Resources. A *Proceedings of the Workshop on Computational approaches to Figurative Language* (pp. 13 - 20).

Lakoff, G. & Johnson, M. (2003). *Metaphors We Live By* (2a ed.). Estats Units: University Of Chicago Press.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. Recuperat des de https://books.google.es/books?id=YiFDxbEX3SUC

Martí, M. A., Taulé, M., Kovatchev, V. & Salamó, M. (2019). DISCOver: DIStributional approach based on syntactic dependencies for discovering COnstructions. A *Corpus Linguistics and Linguistic Theory* (pp. 1 - 33). doi:doi:10.1515/cllt-2018-0028

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, *3*(4), 235 - 244.

Mykowiecka, A., Marciniak, M. & Wawer, A. (2018). Literal, Metphorical or Both? Detecting Metaphoricity in Isolated Adjective-Noun Phrases. A *Proceedings of the Workshop on Figurative Language Processing* (pp. 27 - 33).

Padró, L., Reese, S., Agirre, E. & Soroa, A. (2010). Semantic Services in FreeLing 2.1: WordNet and UKB. A P. Bhattacharyya, C. Fellbaum & P. Vossen (Ed.), *Principles, Construction, and Application of Multilingual Wordnets* (pp. 99 - 105). Global Wordnet Conference 2010. Mumbai, India: Narosa Publishing House.

Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. A *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. ELRA, Istanbul, Turkey.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825 - 2830.

Pennington, J., Socher, R. & Manning, C. (2014). Glove: Global vectors for word representation. A *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532 - 1543).

Pla, X. (1997). *Josep Pla: ficció autobiogràfica i veritat literària*. Quaderns crema.

Rehurek, R. & Sojka, P. (2011). Gensimstatistical semantics in python. *Statistical Semantics; GenSim; Python; LDA; SVD*.

Schierle, M., Schulz, S. & Ackermann, M. (2008). From Spelling Correction to Text Cleaning – Using Context Information. A C. Preisach, H. Burkhardt, L. Schmidt-Thieme & R. Decker (Ed.), *Data Analysis, Machine Learning and Applications* (pp. 397 - 404). Berlin, Heidelberg: Springer Berlin Heidelberg.

Schulder, M. & Hovy, E. H. (2014). Metaphor Detection through Term Relevance. A *Proceedings of the Second Workshop on Metaphor in NLP* (pp. 18 - 26).

Shutova, E., Sun, L. & Korhonen, A. (2010). Metaphor Identification Using Verb and Noun Clustering. A *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 1002 - 1010).

Steen, G. J. (2007). *Finding metaphor in grammar and usage: A methodological analysis of theory and research* (M. Verspoor & W. Spooren, Ed.). Converging Evidence in Language and Communication Research. John Benjamins Publishing.

Strzalkowski, T., Broadwell, G. A., Taylor, S., Feldman, L., Shaikh, S., Liu, T., ... Cases, I. et al. (2013). Robust extraction of metaphor from novel data. A *Proceedings of the First Workshop on Metaphor in NLP* (pp. 67 - 76).

Wierzbicka, A. (1986). Metaphors linguists live by: Lakoff & Johnson contra Aristotle. *Research on Language & Social Interaction*, *19*(2), 287 - 313.