Grau en Lingüística

Facultat de Filologia

Aproximacions quantitatives al significat

Treball de Final de Grau

Aina Garí Soler
Directores: M. Antònia Martí i Mariona Taulé
Barcelona, juny 2015

Aproximacions quantitatives al significat

Aina Garí Soler

Juny 2015

Resum

Aquest treball constitueix una síntesi crítica de l'estat de l'art d'aproximacions quantitatives al significat basades en la Hipòtesi Distribucional de Harris, concretament, de la detecció estadística de *Multiword Expressions* i de la representació del significat mitjançant Models d'Espais Vectorial (VSMs). En veurem el fonament, les característiques i les limitacions actuals principals. Es presenta, també, un experiment dut a terme pel grup d'investigació CLiC que combina l'objectiu de detectar construccions amb el mètode dels VSMs. En aquest marc, i per tal de millorar la representació del significat mitjançant VSMs, faig una proposta per poder-hi representar la polisèmia, una limitació que actualment tenen aquests models.

Abstract

This work constitutes a critical overview of the state of the art of quantitative approaches to meaning based on Harris' Distributional Hypothesis, specifically, of statistical detection of Multiword Expressions (MWEs) and representation of meaning by means of Vector Space Models (VSMs). We will take a look at their basis, characteristics and current main limitations. An experiment carried out by the CLiC research group, which combines the objective of detecting constructions with a metodology based on VSMs, is also presented. Finally, I present a proposal to model polisemy within the VSMs framework, as polisemy poses a still unresolved challenge to these models.

Agraïments

A la Toni i a la Mariona, dues professores, investigadores i directores implicadíssimes que sempre m'han ofert la seva ajuda i el seu temps, que m'han fet sentir molt a gust, i que m'han ensenyat fins i tot més del que es pensen.

Al Ricardo i a la Maria, per escoltar i pensar en les meves propostes, i al Ricardo per implementar-les.

A en Francesc i a l'Àlex, per escoltar-me, orientar-me i fer-me suggerències.

A tots aquells professors i alumnes de Lingüística que van tenir la paciència d'escoltar-me assajar i fer l'exposició d'aquest treball, però en especial als meus pares, per aquesta i moltes altres paciències que tenen.

A tu o a vostè, lector d'aquest treball, amb el desig que la lectura sigui agradable, interessant, instructiva i, tant de bo, inspiradora.

Índex

0 Motivació

Presento aquest treball amb un doble objectiu. Per una banda, la vocació, interès i curiositat personals per la Lingüística Computacional, la Semàntica i les Matemàtiques fan que senti una atracció per aquest tema. La detecció de construccions per mitjans estadístics i l'ús de Models d'Espais Vectorial per representar el significat constitueixen dues noves aproximacions quantitatives a la semàntica que obren línies de recerca que desperten el meu interès. Fer-ne una síntesi esdevé per a mi una manera d'introduir-m'hi, alhora que em prepara el camí per seguir la meva formació futura en Processament del Llenguatge Natural en els estudis de Màster.

Per altra banda, aquest treball espera poder servir com a material de referència preparat, adaptat i dirigit a estudiants de Lingüística interessats també en aquest tema.

1 Introducció

El 1954, el lingüista Zellig Harris proposa la Hipòtesi Distribucional, segons la qual la similitud semàntica de dues expressions lingüístiques està en funció del nombre de contextos que comparteixen. És a dir, si dues paraules apareixen en els mateixos contextos, tindran un significat semblant. Aquesta hipòtesi, proposada des de l'estructuralisme americà, es reforça amb propostes provinents d'altres tradicions. El 1953, el filòsof Ludwig Wittgenstein proposa que el significat d'una paraula està determinat per les seves situacions d'ús. En la mateixa línia, però des de la perspectiva de l'àmbit de l'aprenentatge de segones llengües, Firth (1957, 179) postula que "*You shall know a word by the company it keeps*". Finalment, des de la psicologia, Miller i Charles (1991) aposten també per una aproximació basada en l'ús. Aquests autors consideren que conèixer una paraula és saber com utilitzar-la en el discurs, i fan la suposició que les persones aprenen a utilitzar les paraules observant com aquestes s'utilitzen.

Amb objectius i des de disciplines molt diferents, però amb conclusions molt semblants, aquestes quatre visions sobre el significat i el context són els actuals referents d'un nou enfocament teòric i formal en Semàntica Lingüística que es coneix amb el nom de Semàntica Distribucional.

La Semàntica Distribucional té com a punt de partida la noció de similitud semàntica presentada en la Hipòtesi Distribucional i assumeix que hi ha una relació entre el significat d'una paraula i la seva distribució. Aquesta relació pot ser causal o no, factor que Lenci (2008) estableix com a distintiu d'una hipòtesi distribucional "dèbil" i una "forta". En la seva versió forta, la hipòtesi distribucional considera que la distribució d'una paraula determina la seva representació cognitiva.

L'aparició de la lingüística generativa durant la segona meitat dels anys 50, però, desvia l'atenció que hi havia per l'anàlisi de dades d'ús i pel mètode distribucional. A més, la falta de recursos tecnològics necessaris per dur a terme anàlisis de la distribució amb grans quantitats de dades fa que el mètode distribucional s'apliqui fonamentalment a la descripció de fets lingüístics, per exemple, en l'àrea de la tipologia lingüística. És durant els anys 90 que, amb la progressiva disponibilitat de grans quantitats de dades textuals en format digital, es recuperen els models matemàtics per al processament del llenguatge. En l'àrea de les tecnologies de la llengua, la primera disciplina que aplica mètodes estadístics és el reconeixement de veu, que crea models de parla a partir de corpus orals especialment dissenyats. Posteriorment, els models matemàtics i estadístics s'empren en les tecnologies del text, principalment en Traducció Automàtica. Més recentment, s'està aplicant aquest nou enfocament metodològic en dues línies: la representació del significat i la detecció de *Multiword Expressions*. A partir d'ara, aquests dos temes seran el meu focus d'atenció.

La representació lingüística del significat lèxic, des de la proposta de Katz i

Fodor (1963), s'ha basat en l'ús de primitives semàntiques. Un dels principals problemes d'aquest enfocament, però, i que resta sense resoldre, és l'establiment d'una llista tancada de primitives que gaudeixi d'un ampli consens. Recentment, i en contrast amb aquesta proposta, reprenent la idea de l'anàlisi distribucional, s'ha desenvolupat el que s'anomena *Distributional Models of Meaning*, que modelitzen el significat de les paraules mitjançant vectors construïts a partir dels contextos en què una paraula apareix en un corpus. Aquest és un dels temes que tractarem en aquest treball.

Des de la perspectiva del Processament del Llenguatge Natural, un problema clau és la detecció de *Multiword Expressions*, és a dir, la detecció de grups de paraules que funcionen juntes. Mitjançant diferents tipus de tècniques computacionals basades en l'estadística, es proposen mètodes de detecció de *Multiword Expressions*. Aquestes expressions constitueixen un tema important tant en lingüística cognitiva (Croft i Cruse, 2004) com en lingüística computacional. Es tracta d'unitats de naturalesa molt diferent que no segueixen el principi de composicionalitat i que per tant no han de ser considerades a partir de les seves parts, sinó com a unitats complexes amb un significat únic. Aquesta característica planteja un problema important per al seu tractament computacional. L'altre tema en què centro aquest treball és la detecció automàtica de *Multiword Expressions* per mitjans estadístics.

Referint-se a les col·locacions, un tipus concret de *Multiword Expressions*, Firth (1957) proposa que el significat i l'ús d'una paraula es poden caracteritzar a partir de les paraules amb què coapareix típicament, és a dir, a partir dels seus contextos més habituals. L'ús del context lingüístic és el punt de contacte entre els dos temes centrals d'aquest treball.

A continuació presento l'organització de la memòria. En l'apartat 2 es presenten les *Multiword Expressions* i s'exposen característiques i detalls del procés de la seva detecció automàtica per mitjans estadístics (secció 2.2.1) o per altres mitjans (secció 2.2.2). En l'apartat 3 es fa una introducció sintetitzada als Models d'Espai Vectorial per a la representació del significat. A continuació, en l'apartat 4, s'explica un experiment que combina aspectes de les aproximacions vistes als apartats 2 i 3. En el cinquè apartat presento una proposta per al tractament de la polisèmia en el marc dels Models d'Espai Vectorial. Finalment, en l'últim apartat, veurem algunes limitacions dels models presentats i quina és, per tant, la feina pendent.

2 Les *Multiword Expressions*

2.1 El concepte de *Multiword Expression*

Per qüestions expositives, utilitzaré el terme *Multiword Expression*[1] per referir-me en general a combinacions de dues o més paraules que funcionen com un tot. He escollit aquesta denominació per ser la més utilitzada en Processament del Llenguatge Natural (PLN).

Les *Multiword Expressions* (MWE), o Expressions Multi-Paraula, van molt lligades al concepte més ampli de construcció de la gramàtica cognitiva, que les considera com unitats bàsiques de la gramàtica de les llengües. Hi ha hagut diversos intents de delimitar el concepte i crear-ne una classificació (Croft i Cruse, 2004; Wray i Perkins, 2000; Nunberg et al., 1994; Fillmore et al., 1988), tenint en compte diferents criteris basats, per exemple, en la forma, la convencionalitat, el grau de composicionalitat, la irregularitat sintàctica, el grau de fixació o la freqüència. Això ha donat lloc a diferents acotacions i nomenclatures del concepte: col·locacions, idiomatismes, frases fetes, entre d'altres.

Com observen Fillmore et al. (1988), tots aquests criteris són graduals, la qual cosa suggereix que és més adient tractar les MWEs com un fenomen també gradual, de manera que les podríem situar en un *continuum* que aniria, per exemple, d'expressions completament substantives o plenes lèxicament (*fer el mandra*) a altres de completament formals (*fer X*), passant per expressions que situaríem a mig camí (*fer un petó a X*); o d'expressions gens composicionals (*estirar la pota*) a d'altres de més composicionals (*vermell pujat*) o de completament composicionals (*pilota de futbol*).

Atesa la multiplicitat de criteris per caracteritzar aquestes unitats des de la teoria lingüística i la varietat de classificacions possibles, des del PLN s'ha optat pel terme genèric de MWEs per referir-se a totes elles.

Les MWEs constitueixen un repte de difícil solució en PLN en tasques tan diverses com la Traducció Automàtica, la Recuperació d'Informació o l'anàlisi sintàctica. Per exemple, Baldwin et al. (2004) constaten que el 39% dels errors d'un analitzador sintàctic tenen lloc en MWEs, i Wehrli (2014) mostra com el coneixement de col·locacions té un impacte positiu en el procés d'anàlisi sintàctica.

És per tot això que des de fa ja temps es desenvolupen tècniques i mètodes per identificar les MWEs de manera automàtica o semiautomàtica, aprofitant l'avantatge que proporciona la disponibilitat de corpus de grans dimensions. Al mateix temps, l'obtenció de dades sobre les MWEs ens pot permetre fer-ne una anàlisi lingüística més acurada, a més de facilitar el seu processament, millorant el rendiment de moltes tasques de PLN i fent possible, també, l'objectiu lexicogràfic de llistar-les per a llengües diferents.

[1] He decidit fer servir el nom en anglès ja que és molt usada la seva versió en sigles (MWE).

Evert (2008) proposa distingir dues aproximacions complementàries a l'hora de tractar les MWEs: la teòrica i l'empírica. Des del punt de vista teòric, lingüísticament més interessant, les MWEs són aquelles unitats que els parlants perceben com un tot, és a dir, com una unitat lexicalitzada i no composicional. Des del punt de vista empíric, en canvi, i de manera més general, les MWEs es conceben com combinacions recurrents i predictibles d'elements que mantenen entre ells una atracció que, a partir de mitjans estadístics i dades lingüístiques, es pot quantificar.

És cert que les dues nocions comparteixen alguns aspectes: les combinacions lexicalitzades de paraules tindran probablement una alta freqüència de coaparició, i gran part de les MWEs que trobem estadísticament en un corpus s'explicaran a partir de la seva no-composicionalitat. No hi ha, però, un solapament total entre elles: l'atracció estadística entre paraules no implica no-composicionalitat, i a l'inrevés.

A partir d'ara, en aquest treball utilitzarem l'expressió MWE en el seu sentit més empíric, és a dir, referint-nos de manera general a una combinació de paraules que tendeixen a coocórrer (a aparèixer juntes) en el discurs (Evert, 2008).

2.2 Detecció de MWEs

Entre els diferents mètodes per a la detecció de MWEs podem distingir, típicament, entre aquells que busquen un tipus en concret de MWE, com per exemple, MWEs que continguin una paraula determinada o que tinguin un determinat patró categorial, com ara preposició-nom i aquells que fan una cerca general per descobrir tot tipus de MWEs. Independentment de l'objectiu de la recerca, el mètode utilitzat pot ser més o menys complex i pot anar de la detecció a ull nu fins a l'aplicació de tècniques estadístiques sobre text pla o anotat.

En els propers apartats es presenten algunes de les tècniques estadístiques que es poden utilitzar en la detecció d'aquestes unitats i posteriorment veurem el cas de la base de coneixement StringNet, la construcció de la qual no requereix de mitjans estadístics.

2.2.1 Aproximació estadística a les MWEs

L'aproximació estadística a la detecció de MWEs té la seva base en l'enfocament empíric d'aquest tipus d'unitats. En aquesta aproximació, destacaré els treballs d'Evert (2008) i Pecina (2010), ja que constitueixen les propostes més avançades en aquest tema. Tots dos autors se centren en un tipus específic de MWE, les col·locacions. Les col·locacions són combinacions, habitualment, de dues paraules de base lèxica amb significat semi-compositiu, com ara *pluja torrencial, fer un petó*. Evert (2008) presenta un mètode per identificar col·locacions a partir de

l'aplicació de mesures estadístiques basades en la freqüència de les paraules. Al seu torn, Pecina (2010) desenvolupa un mètode d'extracció de col·locacions mitjançant la combinació de diferents mesures estadístiques, i posteriorment fa una avaluació del seu mètode.

Sota la mirada de l'estadística, les paraules d'un corpus no són més que elements desprovistos de tot significat que es distribueixen entre ells d'una manera concreta que es pot copsar matemàticament. Una col·locació, llavors, és un parell de paraules que coapareixen més freqüentment que el que s'esperaria per atzar.

A continuació presentarem els conceptes que considerem fonamentals per entendre les bases d'aquest enfocament estadístic. Després es discutirà sobre algunes decisions de caire lingüístic que s'han de prendre en el procés, com ara la definició de context o l'aplicació de filtres. Finalment comentarem algunes de les mesures estadístiques més utilitzades i també maneres de presentar i avaluar els resultats.

1. Conceptes bàsics

Freqüència

Cal distingir dos conceptes bàsics de freqüència que s'utilitzen sovint sota el mateix nom però que no són equivalents i que es basen en la distinció entre valors absoluts i relatius.

Freqüència absoluta. És el nombre de vegades que apareix algun element en un conjunt de dades textuals. Per exemple, podem dir que la paraula *accident* apareix 10 vegades en un corpus donat. Cal fixar-se, però, que aquest valor aïllat és insuficient per saber si la freqüència és alta o baixa: depèn de les dimensions del corpus. Per poder utilitzar aquesta dada, doncs, cal tenir en compte el nombre total de paraules del corpus.

Freqüència relativa. Si sumem les freqüències absolutes de totes les paraules, obtenim el nombre total de paraules del corpus. Aquesta és una dada fonamental per entendre el pes que té cada paraula en el corpus. Dividint la freqüència absoluta de cada element entre el nombre d'elements total, obtenim la freqüència relativa, que ens donarà informació, normalment en forma de tant per u, de la presència de la paraula en el corpus. Així, la freqüència de la paraula *accident*, de valor absolut 10, tindrà molta rellevància si es troba en un corpus de 1000 paraules, però en tindrà poca si el tamany del corpus és d'un milió de paraules.

Dos conceptes més, també relacionats amb la freqüència però independents dels dos anteriors, són d'utilitat per a l'anàlisi estadística de les MWEs:

Freqüència de coaparició

Freqüència de coaparició observada. És el nombre (absolut) de vegades que dues paraules coapareixen, és a dir, que comparteixen context. La definició de coaparició es discutirà amb detall al punt 2.

Freqüència de coaparició esperada. El concepte de valor esperat és molt important en estadística i serà clau per dur a terme el càlcul de mesures d'associació. La freqüència de coaparició esperada és la freqüència amb què esperaríem que dues paraules coapareguessin si les paraules del corpus estiguessin ordenades de manera **aleatòria**. Comparant estadísticament els valors observat i esperat, si aquests difereixen de manera significativa, podem determinar si dues paraules tenen o no força d'atracció entre elles. Les dues freqüències es poden comparar mitjançant *mesures d'associació*. Veurem, però, que l'ús d'aquests valors no és suficient per a la detecció de MWEs.

Mesures d'associació

Són mesures o fórmules matemàtiques que quantifiquen l'atracció entre dues paraules o la seva associació estadística. N'hi ha moltes[2] i no s'ha provat que n'hi hagi una que sigui millor que les altres. Per a cada parell de paraules, una mesura d'associació ens dóna una puntuació (*score*) que s'interpreta d'una manera o d'una altra en funció de la mesura utilitzada. A més puntuació, més atracció.

Hipòtesi nul·la

En estadística es treballa amb l'anomenada hipòtesi nul·la, que suposa, en el nostre context, que les dues paraules que ens ocupen no tenen atracció entre elles. Aquest seria el cas de l'aleatorietat, és a dir, de la freqüència esperada. Si la freqüència observada i l'esperada són semblants, la hipòtesi nul·la serà certa i no es podrà considerar que hi hagi atracció. Si, en canvi, l'atracció entre les dues paraules és significativa, és a dir, prou forta com per descartar la hipòtesi nul·la, llavors considerarem que les dues paraules formen un candidat a col·locació. Moltes de les mesures d'associació que es fan servir tenen aquest concepte de base.

Precisió i cobertura

Es tracta de dues mesures importants en classificació binària. En un conjunt d'elements, es poden definir classes per agrupar-los. N'hi ha que són d'un tipus o d'un altre: A o B. En el nostre cas, de tots els parells de paraules, n'hi ha que són col·locacions (A) i n'hi ha que no (B). El nostre objectiu és trobar automàticament quins pertanyen a A i quins a B. En aquest procés, però, es poden produir dos tipus d'errors:

[2] Pecina (2010) n'aplica i avalua 82.

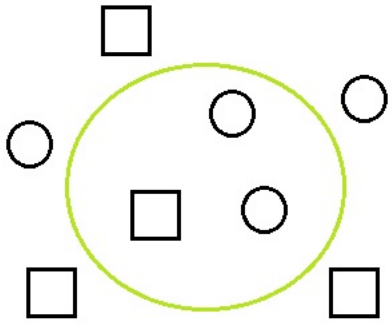

Figura 1: Precisió i cobertura

- Considerar com a col·locacions parells de paraules que no ho són, és a dir, falsos positius;
- No considerar com a col·locacions parells de paraules que sí que ho són, és a dir, falsos negatius.

La **precisió** és la proporció de col·locacions que han estat detectades correctament respecte de tots els elements que s'han trobat com a col·locacions. Dit d'una altra manera: d'aquells que s'han identificat, quants són correctes. Si la precisió és molt alta, significa que hi ha pocs o cap fals positiu.

La **cobertura** (o *recall*, en anglès) és la proporció de col·locacions que han estat detectades correctament respecte del total de col·locacions reals que hi havia. És a dir: d'aquells que s'havien d'identificar, quants se n'han identificat. Valors elevats de cobertura són equivalents a tenir pocs falsos negatius.

En la Figura 1 es mostra un senzill exemple il·lustrat d'aquestes dues mesures. En aquesta figura, la circumferència gran envolta aquells elements que han estat considerats com a correctes. Els quadrats són els incorrectes i les circumferències, els correctes. La precisió (p) seria del 66,6%, ja que 2/3 dels elements que s'han agafat són correctes. La cobertura (c), en canvi, seria del 50%, ja que només s'han agafat 2 dels 4 elements que eren correctes (2/4). En tant per u, p=0,66 i c=0,5.

Distribucions i Llei de Zipf

En estadística i teoria de la probabilitat es parla de **distribucions de probabilitat** d'una variable per referir-se a una funció que codifica o aproxima les probabilitats de cada esdeveniment. Gràficament, es poden col·locar les freqüències dels diferents esdeveniments (en el nostre cas, diferents parells de mots) en un eix de coordenades, i la manera com aquestes freqüències estan repartides dóna lloc a una distribució o a una altra. Una de les distribucions més àmpliament utilitzades

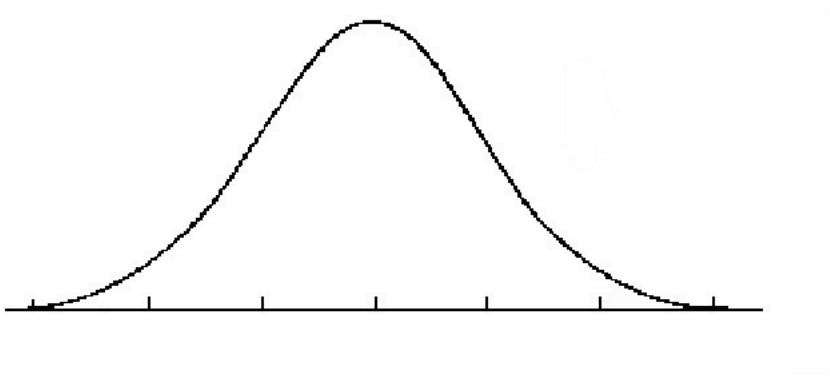

Figura 2: Distribució normal

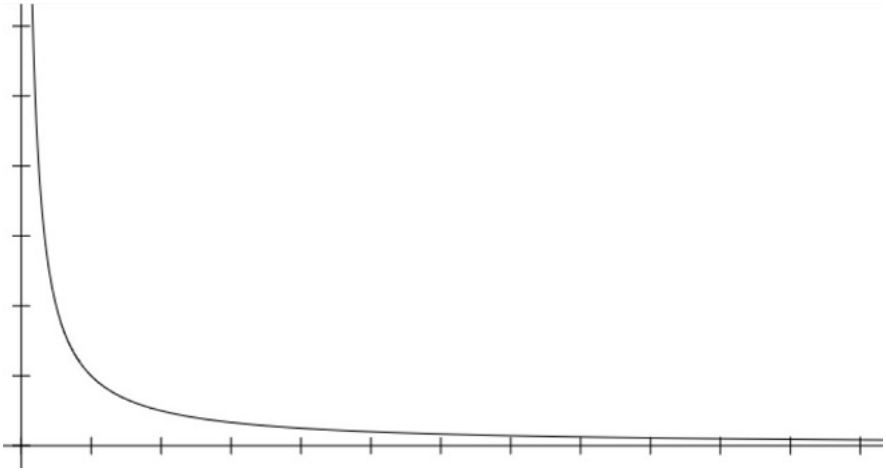

Figura 3: Distribució Zipfiana

és l'anomenada **distribució normal**, representada en la Figura 2. Aquesta distribució es caracteritza per una majoria de valors intermedis, amb valors extrems menys freqüents. És important perquè s'ha observat que molts fenòmens naturals s'adapten a aquesta corba: per exemple, l'alçada de les persones.

Una altra distribució, de gran rellevància en lingüística, és la **distribució Zipfiana** (Figura 3). S'ha observat que el llenguatge segueix una distribució en què un nombre relativament reduït de paraules tenen alta freqüència i són majoria les paraules amb freqüències baixes.

2. Definició de la coaparició

Abans d'entrar en el tema de com detectar MWEs cal definir el concepte de coapa-

rició, també anomenada coocurrència: Què vol dir que dues paraules coapareixen? Per dur a terme recomptes de les vegades en què dues paraules coapareixen, cal saber en quines circumstàncies es pot dir que això passa, és a dir, en quin moment es considera que les paraules comparteixen context. Aquesta és una decisió que correspon al lingüista i de la qual dependrà el resultat final. En general, s'estableixen tres grans tipus de coaparició possibles:

Coaparició superficial

Es basa en la proximitat espacial de les paraules en la cadena textual. Segons aquest tipus de coaparició, dues paraules coapareixen si apareixen pròximes l'una a l'altra. Una de les característiques principals de la coaparició superficial és que no requereix una anàlisi sintàctica del corpus. Es critica, però, la tria arbitrària dels paràmetres i el fet que pot no ser adequada per a llengües amb un ordre relativament lliure o estrictament fixat.

Si tenim en compte aquest tipus de coaparició cal decidir sobre una sèrie de qüestions:

Com determinar la proximitat. L'investigador ha de determinar lliurement l'abast del context, és a dir, el nombre de paraules que es tindran en compte. Aquest espai s'anomena **finestra de paraules**. Per exemple, si considerem que una paraula és pròxima a una altra quan apareix no més enllà de 2 paraules a dreta i esquerra, tindríem una finestra simètrica de 2 paraules a l'esquerra i 2 a la dreta (L2,R2)[3]. També és freqüent tenir en compte paraules directament adjacents (bigrames (L0,R1)). Aquest criteri és molt variable i se sol decidir de manera arbitrària o segons la tasca que es dugui a terme.

Puntuació, caràcters no alfabètics, *stop-words*, MWEs conegudes. El lingüista ha de prendre decisions sobre diferents aspectes relatius a les propietats dels textos escrits. En primer lloc, ha de decidir si els signes de puntuació o els números seran tractats com una paraula més o no. Rebutjar-los significa que no es comptaran en la finestra de paraules i, per tant, no formaran part de cap candidat a MWE ni augmentaran la distància entre dues paraules. Depenent del tipus de MWE que es busqui, per exemple, una en què tots els elements siguin de base lèxica, és possible que interessi també ignorar totes les paraules gramaticals (o *stop words*). També és important decidir si algunes MWEs ja conegudes i determinades seran tractades com a una sola paraula o com a més d'una. Aquest és el cas, sobretot, de les que funcionen com a paraules gramaticals, com ara *no obstant això, com que....*

Límits d'oració. Una altra decisió que cal prendre és si les finestres de paraules haurien de definir-se ignorant o no els límits de les oracions. Per això cal tenir en

[3]L i R vénen de l'anglès *left* i *right*, respectivament.

compte com està constituït el corpus[4].

Coaparició textual

En aquest cas, es considera que dues paraules coapareixen si apareixen en la mateixa unitat textual. La unitat textual pot ser una oració, un sintagma o un document sencer. Aquesta definició del context permet copsar coaparicions que s'escapen de l'abast de la coaparició superficial, ja que en general té en compte unitats d'extensió més llarga.

La coaparició textual és més fàcil d'implementar que la coaparició superficial, ja que té menys paràmetres per determinar, i és més robusta: troba més coaparicions. Té, però, l'inconvenient que les grans quantitats de dades que genera poden ser molt costoses de processar.

Coaparició sintàctica

Es considera que dues paraules coapareixen sintàcticament si guarden entre elles una relació sintàctica. En aquest cas, és imprescindible disposar d'un corpus analitzat sintàcticament. Els diferents tipus de relacions sintàctiques se solen tractar per separat i es pot decidir quins tipus de dependències interessen i quines no es tindran en compte. A diferència de la coaparició superficial, la sintàctica pot identificar moltes relacions implícites, no observables directament, i a llarga distància, de manera que l'ordre relativament lliure dels elements a l'oració d'una llengua no afectarà els resultats.

3. Extracció de parells i freqüències

Un cop definida la coaparició, el següent pas en el procés de detecció de MWEs consisteix en l'extracció de parells de paraules i la informació sobre la seva freqüència. En funció del tipus de context definit, s'obté de manera automàtica una llista de parells de paraules a partir del text. Per exemple, si la paraula *té* ha coaparegut amb *lloc*, afegim el parell (té, lloc) a la llista. En aquest moment cal decidir si es **lematitzarà**, cas en el qual parells com (té, lloc) o (va tenir, lloc) es convertirien tots en un element lematitzat (tenir, lloc), que agruparia totes les seves possibles variants amb els respectius recomptes de freqüència.

Per a cada parell de paraules cal aportar quatre dades, que s'obtenen a partir del corpus:

[4]Si el corpus és una recopilació de molts textos relativament breus i de temes molt variats (La Viquipèdia, per exemple), possiblement l'opció més adequada seria posar barreres en els límits d'oració, ja que d'una altra manera sovint estaríem comptant com a coocurrents paraules que originalment formaven part de textos diferents. En canvi, però, si el corpus és, per exemple, una única obra literària, segurament no sigui necessari aturar-nos en els canvis d'oració

- La freqüència de coaparició, és a dir, el nombre de vegades que les dues paraules han compartit context.
- Les freqüències individuals de cada una de les dues paraules en el corpus.
- El nombre total de tokens, que depèn del tipus de coocurrència definit: en la coocurència superficial serà el nombre de tokens que poden formar part de les finestres; a la coaparició textual, el nombre d'unitats en què es basi (normalment frases); i en la sintàctica és el nombre total de parells que s'han trobat per a una mateixa relació.

4. Filtres

L'últim pas que cal realitzar abans de procedir amb els càlculs sobre el grau d'atracció entre paraules és l'aplicació opcional d'un filtre amb l'objectiu de reduir la quantitat de dades que s'obtenen en el procés d'extracció. Podem aplicar-ne de dos tipus: filtre de freqüència i filtre de categoria.

Filtre de freqüència

Consisteix a eliminar de la llista de parells de paraules aquells que tinguin un nombre d'aparicions inferior a un llindar (*threshold*) que s'ha d'establir. Aquest filtre, si bé simplifica notablement els càlculs, presenta problemes:

En primer lloc, si s'accepta la distribució *Zipfiana* de les paraules, la majoria d'elles apareixen poques vegades. Amb aquest filtre es retallaria un volum considerable de dades, perdent d'aquesta manera una informació valuosa i no detectable per altres mitjans. Es restringiria el càlcul a les paraules més comunes, que són probablement aquelles de les quals ja en sabem més o en podem obtenir informació més fàcilment. S'estaria limitant, per tant, els nostres resultats.

En segon lloc, el filtre per freqüència es duu a terme molts cops per millorar els resultats de mesures estadístiques que assumeixen una distribució normal de les observacions. Aquest tipus de mesures tenen una important limitació a l'hora d'analitzar coaparicions "inusuals", rares, o de baixa freqüència (Dunning, 1993; Moore, 2004). No solen ser fiables, però, si s'apliquen a corpus, ja que el llenguatge no segueix una distribució normal.

Filtre per categoria

Hi ha certes combinacions de categories que no poden formar mai una MWE: és el cas, per exemple, de conjunció-preposició. Per tant, es poden excloure i així reduir càlculs en principi innecessaris.

5. Càlcul i aplicació de mesures d'associació

Què són

Les mesures d'associació són fórmules que assignen una puntuació a cada parell de paraules segons el grau d'atracció que aquestes presenten entre elles. Hi ha moltes mesures diferents i amb bases matemàtiques molt variades. Pecina (2010) computa i avalua els resultats de fins a 82 mesures diferents, algunes de les quals, però, no són pròpiament d'associació estadística, sinó d'anàlisi del context, i no es comentaran en aquest treball.

Què es necessita

Per fer servir algunes de les mesures d'associació (les més senzilles), cal saber només les freqüències de coaparició observada i esperada. Aquesta última es pot obtenir a partir de les freqüències individuals de cada paraula i del nombre total de paraules del corpus.

Altres mesures estadístiques, més complexes, requeriran que per a cada parell es computin diverses freqüències. En aquest cas, per a cada parell de paraules (w1,w2)[5], cal obtenir una taula de contingència amb la següent informació sobre els valors observats:

- freqüència de coaparició observada;
- unitats en què apareix w1 però no w2;
- unitats en què apareix w2 però no w1;
- unitats en què no apareix ni w1 ni w2.

La definició d'unitat depèn del tipus de coaparició que s'hagi definit: les finestres de la paraula en qüestió en el cas de la superficial, la unitat escollida en el cas de la coaparició textual (per exemple, la frase) i els parells de paraules en el cas de la coaparició sintàctica.

Per a algunes mesures caldrà calcular també una taula equivalent amb els corresponents valors esperats. A continuació s'expliquen tres mesures d'associació.

Algunes mesures d'exemple

Pointwise Mutual Information (PMI). Aquesta mesura d'associació prové de la teoria de la informació i serveix per medir, en bits, la "informació mútua" o compartida de les dues paraules. Si dues paraules no comparteixen informació (no se "solapen"), es diu que són independents l'una de l'altra, és a dir, no estan relacionades i, per tant, no s'"atrauen". En aquest cas, conèixer una de les paraules no ens aporta cap informació sobre l'altra: el valor de la PMI seria 0.

[5] *w* és l'inicial de *word*, l'anglès per *paraula*.

El test de Pearson de la X^2. El test de Pearson de la X^2, o de "la bondat d'ajustament", consisteix a comparar la distribució observada de les paraules amb una distribució hipotètica (que assumeix la independència o la no-atracció entre dues paraules basant-se en la hipòtesi nul·la) i quantificar-ne la discrepància per veure fins a quin punt el desajustament entre les dues distribucions, l'esperada i l'observada, és degut a l'atzar o no. Aquesta mesura no dóna bons resultats si s'aplica a quantitats de dades no gaire grans. El *test exacte de Fisher* seria més adequat en aquestes circumstàncies, però té l'inconvenient de ser molt costós computacionalment. Per això existeix també una versió corregida del test de la X^2 que contrarresta parcialment el biaix, la *correcció de Yates*.

Log-Likelihood ratio. Aquest test estadístic serveix per mesurar l'ajustament o la correspondència entre allò esperat i allò observat. En concret, compta quantes vegades un dels valors (l'esperat i l'observat) supera l'altre. El seu valor és més directament interpretable que el de la X^2.

Combinació de mesures

Una possibilitat és combinar dues o més mesures per a la detecció de MWEs. Així, Pecina (2010) calcula les puntuacions de 82 mesures i en crea una de nova a partir de la combinació de totes elles, millorant notablement els resultats. A continuació, proposa un algoritme que redueix aquesta combinació de 82 mesures a 13, eliminant aquelles que no contribueixen al resultat final o que, fins i tot, l'empitjoren. Amb aquesta tècnica aconsegueix reduir àmpliament la complexitat de la mesura sense gairebé perjudicar-ne el resultat.

En general, no es pot dir que hi hagi una mesura d'associació que sigui millor que les altres. Els resultats de les mesures seran diferents, millors o pitjors en funció de l'objectiu, de la llengua en qüestió, de la definició de coaparició, dels llindars que s'hagin establert o del corpus amb què es treballi (segons el seu tamany i composició). En definitiva, diferents mesures obtenen diferents resultats.

Les mesures presentades són del tipus que s'anomenen *no paramètriques*, és a dir, serveixen per a dades que s'ajusten a qualsevol distribució. Altres mesures que es fan servir, com per exemple la *z-score*, requereixen un ajustament a la distribució normal que les dades no compleixen. Aquest desajustament provoca biaixos en els resultats i en teoria invalida qualsevol argument matemàtic per utilitzar-les. Tot i així, val la pena provar-les, ja que poden donar resultats interessants.

6. Rànquing o llindar

Un cop s'han fet els càlculs i es disposa de les puntuacions de tots els parells de paraules amb una determinada mesura d'associació, cal decidir com es determinarà quins són candidats a col·locació. Una manera de fer-ho és mitjançant un

threshold, és a dir, establint un llindar mínim de manera que només aquells parells de paraules que el superin siguin considerats candidats. La decisió del llindar és arbitrària.

Un altre procediment és el de fer un rànquing de tots els resultats ordenant-los de major a menor. En aquest cas es tracta la "col·locativitat" com un fenomen gradual, sense obtenir una llista fixa de candidats. D'aquesta manera no és possible la posterior avaluació, ja que per dur-la a terme, com veurem, es necessita una classificació estricta.

Una tercera opció és fer una combinació de rànquing i llindar: col·locar tots els parells en ordre i seleccionar-ne un nombre determinat començant per dalt, creant així una llista dels n millors.

Finalment, un altre mètode que combina els dos anteriors consistiria a establir tres o més llindars, classificant les paraules en "molt col·locacional", "una mica", "poc", "gens". Un altre cop, però, la tria d'aquests llindars és completament arbitrària i augmentar el nombre de categories augmenta el grau d'error. Tot i així, aquests llindars es podrien establir *a posteriori* un cop feta l'avaluació.

7. Avaluació

Una manera d'avaluar el rendiment de cada mesura és comparar els resultats que s'han obtingut amb judicis humans fets prèviament sobre tots els parells de paraules que coocorren. El procés habitual consisteix en una tasca en què a dues o més persones se'ls dóna la llista de parells obtinguts i aquests els han de classificar d'acord amb unes categories i uns criteris prèviament definits i iguals per a tots els participants. Es pot demanar una tasca de classificació binària (col·locació o no col·locació) o de classificació ordinal en diferents categories (de més a menys col·locacional), però és més senzilla la primera opció. En un procediment similar amb 3 individus, Pecina (2010) calcula el percentatge d'*Inter-annotator agreement* (acord entre anotadors) en la classificació binària, obtenint un valor del 56%, dada que dóna compte de la vaguetat del concepte. És recomanable acceptar com a col·locacions verdaderes només aquells parells que gaudeixin d'un consens ampli o total entre els individus que han participat en la classificació. Posteriorment, es pot comprovar l'adequació dels resultats de cada mesura als judicis humans mitjançant el càlcul de la precisió i la cobertura (introduïts al punt 1). Com més alts siguin aquests dos valors, més propera és la mesura al resultat desitjat. També es pot comparar els diferents valors de precisió i cobertura de diversos llindars d'una sola mesura, per descobrir quin és el llindar òptim en cada cas. S'observa que llindars molt restrictius -és a dir, que accepten menys col·locacions com a vàlides- fan augmentar la precisió (els candidats són correctes) però disminuir la cobertura (ja que es deixen d'acceptar altres parells que també eren correctes). A

mesura que se suavitza el llindar, aquests valors es van intercanviant de manera que amb un llindar molt baix es perd precisió però es guanya cobertura.

2.2.2 Detecció de patrons multiparaula a partir d'n-grames: StringNet

StringNet[6] (Wible i Tsao, 2010) és una base de coneixement consistent en una recopil·lació de patrons lexico-sintàctics de l'anglès. A partir d'un procediment molt senzill i sense necessitat d'aplicar cap mesura estadística obtenen un recurs potencialment útil per a la investigació lingüística. La metodologia que empren és independent de la llengua i es podria, per tant, crear StringNets per a qualsevol llengua, sempre i quan es disposi de corpus anotats amb informació sobre la categoria gramatical de les paraules i el seu lema.

StringNet s'ha construït a partir del *British National Corpus* (BNC). A partir del corpus es deriven **n-grames híbrids** de fins a 8 paraules amb un llindar de freqüència de 5 aparicions. Un n-grama híbrid és un n-grama que pot contenir informació sobre paraules, lemes i/o categories. D'aquesta manera es es reflecteix la propietat de les construccions de ser **substantives** (o lèxicament plenes), com ara *estirar la pota*; o **formals** (o lèxicament obertes), com ara *des del X punt de vista*, on X seria un pronom possessiu variable (*meu*, *teu*, *seu*...).

Com a exemple d'un n-grama híbrid obtingut a partir d'*estira la pota* tenim:

- estirar la pota,
- [verb] [det] pota,
- estira la [nom]

...

Sobre aquestes dades s'aplica un procés de refinament (*pruning*), amb l'objectiu de compactar el nombre de patrons i eliminar redundància. El refinament es duu a terme verticalment i horitzontalment. El refinament vertical eliminaria el segon dels següents n-grames per redundància, ja que [prep] sempre correspon a *de*:

El meu punt de vista

El meu punt [prep] vista

En canvi, el refinament horitzontal compara n-grames de diferent llargada. En el següent exemple, s'eliminaria el més curt:

[dposs] punt de

[6] http://www.lexchecker.org/

[dposs] punt de vista

Com es pot observar, es tracta d'una metodologia simple però que dóna resultats rellevants i crea un recurs de lliure disposició. El problema és que són poques les llengües que disposen d'un corpus de les dimensions del BNC enriquit amb informació morfològica de qualitat.

3 Els *Vector Space Models*

3.1 Una nova manera de representar el significat

L'aproximació tradicional al significat lèxic, basada en primitives semàntiques, presenta diversos inconvenients. Les primitives haurien d'expressar les unitats mínimes del significat en un conjunt tancat i considerablement menor que el nombre de paraules que cal definir, però alhora haurien de ser capaces de representar tots els significats. La realitat és, però, que no s'ha aconseguit una llista de primitives que compleixi aquestes característiques i que gaudeixi d'ampli consens. A més, no s'ha demostrat que les primitives tinguin una base cognitiva vàlida, ja que paraules amb una descomposició en primitives més complexa haurien de ser més difícils de processar, però no és el cas (Fodor et al., 1975). Per a més detalls, es pot consultar Boleda i Erk (2015). Els *Vector Space Models* (VSMs), o Models d'Espai Vectorial, són un sistema de representació de paraules o d'unitats textuals que s'utilitza en Semàntica Distribucional. Com ja s'ha vist, segons la Hipòtesi Distribucional (Harris, 1954), el significat d'una expressió lingüística ve donat pel context lingüístic en què aquesta apareix. Es pot afirmar, llavors, que el significat d'una paraula és la suma dels diferents contextos en què apareix. Aquesta proposta es pot formular en termes de VSMs, una alternativa a les primitives pel que fa a la representació del significat. En aquesta alternativa, els trets subjectius i sense base empírica que proposen les primitives se substitueixen per contextos reals d'ús de la llengua.

La Semàntica Distribucional constitueix una nova manera de tractar la semàntica lèxica i els VSMs proporcionen la metodologia necessària per dur a terme una modelització d'aquestes característiques. En els VSMs, les paraules es defineixen automàticament -i no manualment- a partir dels seus contextos de coaparició, que constitueixen els trets que les caracteritzen. Hi ha, per tant, tants trets possibles com paraules a definir, ja que cada paraula pot ser alhora un context. S'ha comprovat que aquests models recullen molts dels aspectes que, idealment, les primitives haurien de copsar, i són capaços d'apropar-se o fins i tot superar els humans en algunes tasques, com ara la detecció de sinònims, els judicis de similitud, la categorització, la identificació de propietats característiques, etc. (Baroni i Lenci, 2010).

Per tant, els models de Semàntica Distribucional, constitueixen una proposta prometedora per modelar una part important del significat només a partir del context lingüístic. De fet, és molt habitual que endevinem el significat d'una paraula desconeguda a partir del context en què apareix.

Una crítica que reben tant aquests models com els tradicionals és el que es coneix com *the symbol grounding problem* o el problema del fonament del símbol (Harnad, 1990). Aquesta crítica planteja la insuficiència dels models que repre-

	mar	viatge	...
tren	0	4	...
vaixell	5	2	...
barca	4	1	...
...	...	...	...

Taula 1: Exemple de matriu simplificada

senten el significat d'elements lingüístics mitjançant altres elements lingüístics i sense cap altre tipus d'informació d'una altra naturalesa. De fet, hi ha evidències que, cognitivament, els conceptes tenen una base important en la nostra experiència sensoriomotora (Pulvermüller, 2005). És per això que recentment s'està afegint informació visual als models de Semàntica Distribucional a partir d'imatges i tècniques de visió per computador, millorant significativament els resultats de models anteriors que utilitzen informació únicament lingüística (Bruni et al., 2014).

Com que els VSMs infereixen el significat automàticament a partir de dades lingüístiques reals, és imprescindible disposar de corpus de grans dimensions a partir dels quals extreure tota la informació per modelar el significat.

3.2 Aspectes formals

En els VSMs, les paraules (o els lemes) es representen amb vectors i els seus contextos són les coordenades d'aquest vector. Un **vector** és un objecte geomètric amb determinades característiques que podem representar gràficament com una fletxa o numèricament com una llista ordenada de nombres.

Si per a cada paraula del corpus en qüestió obtenim un vector, podem agrupar-los en una taula o **matriu**, que tindrà tantes files i columnes com paraules diferents tingui el corpus. Cada fila és un vector que té codificada numèricament la coaparició de la paraula que dóna lloc a la fila amb tots els contextos possibles, que són a les columnes. A tall d'exemple, podem considerar la matriu simplificada[7] de la Taula 1 que conté les paraules *tren*, *vaixell* i *barca*. En aquest exemple, les coordenades són nombre de coaparicions: *tren* ha coaparegut 0 vegades amb *mar* i 4 amb *viatge*. El seu vector seria $(0, 4, ...)$.

Podem representar les paraules de la nostra matriu en un Espai Vectorial (o Espai Semàntic), com veiem en la Figura 4.

[7]Cal tenir en compte que en els VSMs amb què es treballa les matrius tenen milers de dimensions, de manera que resulta humanament impossible imaginar-les representades gràficament en l'Espai Semàntic (Figura 4).

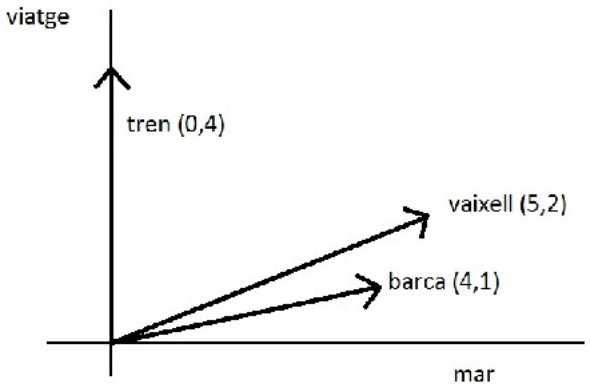

Figura 4: Espai semàntic

En l'Espai Semàntic s'aprecia d'una manera més intuïtiva la relació entre les diferents paraules que ens ocupen. En aquest cas, *mar* i *viatge* són les dimensions, i la resta de paraules estan situades en funció d'aquestes dues. Es pot veure clarament que *tren* no té res a veure amb *mar* i que *barca* té poc a veure amb *viatge*, mentre que *vaixell* se situa al mig, per estar relacionat amb totes dues. Fixem-nos també que, en aquest exemple, *barca* i *vaixell* són més pròximes entre elles que amb *tren*, que queda apartat. Aquesta idea de proximitat en l'espai és la que, com veurem, dóna lloc al concepte de **similitud**.

Aquesta és la idea de base dels VSMs. A partir d'aquí, s'hi poden fer diverses modificacions, afegiments o canvis de paràmetres. Vegem breument algunes de les moltes possibilitats que ens proporcionen els VSMs.

Altres tipus de matrius

A més de matrius paraules-context, es poden fer matrius **paraules-document** o **paraules-patró** (Turney i Pantel, 2010), depenent del nostre objectiu. En el primer cas, paraules-document, es treballa amb la hipòtesi que la freqüència de paraules en un document tendeix a indicar la importància d'un document respecte d'una consulta (Salton et al., 1975) i s'utilitza per a recuperació d'informació, clusterització o classificació de documents, entre d'altres.

Pel que fa a les matrius paraules-patró, en què els vectors són parells de paraules i les columnes, els contextos, es poden utilitzar per a tasques com identificació d'analogies o similitud de patrons. Baroni i Lenci (2010) proposen un model en què tracten aquestes relacions mitjançant tensors de 3r ordre, que són objectes geomètrics amb volum: constitueixen una matriu que a més d'alçada i amplada, té profunditat. De fet, les matrius també es poden anomenar tensors de 2n ordre.

Aquest treball se centra en les matrius paraules-context.

Aspectes lingüístics

En la construcció d'un VSM cal prendre també decisions de caràcter lingüístic, aquelles relacionades amb la **tria del corpus** i a la **definció de context**. Les característiques del corpus dependran de l'objectiu o del tipus de discurs que es vulgui estudiar. Cal remarcar que els VSMs són independents de la llengua. El criteri general a l'hora d'escollir el corpus és que com més gran sigui, més bons seran els resultats (Curran i Moens, 2002), ja que, entre d'altres raons, amb moltes dades els errors que hi pugui haver al corpus tenen menys pes estadístic.

La definició del context és també una decisió clau. Com hem vist al punt 2 de l'apartat 2.2.1, el context es pot definir en termes de coaparició superficial, textual o sintàctica. No n'hi ha una que sigui millor que les altres: cada una dóna diferents resultats i cal comprovar quina modelització dóna millors resultats.

Recomptes i Similitud

Com s'ha comentat, les coordenades dels vectors, a la pràctica, no necessàriament es corresponen amb el nombre de coaparicions, és a dir, amb la freqüència de coaparició absoluta. Normalment, aquests recomptes es processen mitjançant **mesures d'associació** com ara les ja conegudes *(Positive) Pointwise Mutual Information* o *Log-Likelihood Ratio*, o d'altres com, en el cas de les matrius paraules-document, la *tf-idf (term frequency-inverse document frequency)*. Aquestes mesures serveixen perquè la coaparició amb paraules poc freqüents tingui més pes que la coaparició amb paraules més freqüents, ja que probablement és menys informativa i no ens ajudaria a discriminar significats per calcular la similitud. En altres paraules, el fet que *tren* coaparegui amb *viatge* ens dóna més informació sobre el significat de *tren* que el fet que *tren* coocorri amb una paraula molt freqüent, com ara l'article *el*. L'aplicació de mesures d'associació farà que la relació entre *tren* i *viatge* tingui més puntuació, encara que *el tren* aparegui més vegades al corpus. Aquestes puntuacions seran les noves coordenades del vector.

Com ja s'ha dit, en els VSMs es parla de **similitud** en termes de proximitat en l'Espai Semàntic. Així, matemàticament, la similitud s'obté a partir de càlculs de distàncies o d'angles. La mesura més àmpliament utilitzada és el cosinus, que té en compte l'angle que formen dos vectors (és a dir, dues paraules). L'avantatge de fer servir una mesura que té en compte l'angle enlloc de la distància, i la raó per la qual dóna millors resultats, és que no es té en compte la llargada del vector (és a dir, la seva magnitud o puntuació), sinó la tendència a anar cap a una direcció. És a dir, si dues paraules fossin sinònims perfectes[8], però una d'elles fos menys

[8]Entenent com a sinonímia perfecta, en el marc de la Semàntica Distribucional, una coincidència de contextos total i proporcional a la freqüència de cada paraula.

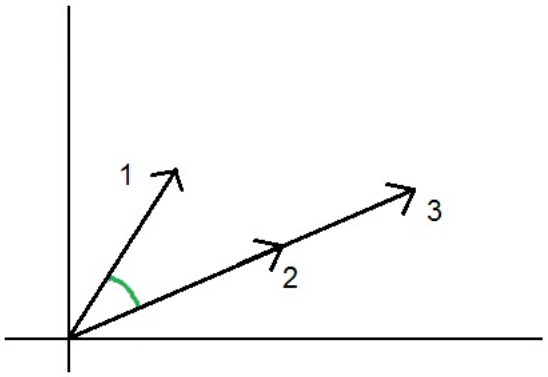

Figura 5: Vectors i els seus angles

freqüent que l'altra, el cosinus, més robust, ignoraria la diferència en freqüència i treuria la conclusió que són completament iguals, mentre que una mesura de distància (com ara la distància Euclídea) donaria importància a aquesta diferència i no ens les presentaria com a sinònimes.

En la Figura 5 es pot apreciar la idea d'angles entre vectors. El vector 1 guarda amb 2 i 3 un angle representat de color verd. Com més gran sigui l'angle, menys similars seran les paraules. Els vectors 2 i 3 tenen entre ells un angle de 0 graus, que equival a la màxima similitud o sinonímia. En canvi, si utilitzéssim una mesura de distància, les paraules serien semànticament distants.

Reducció de dimensions

Una matriu de coaparició que conté les puntuacions de tots els parells de paraules que apareixen en el corpus conté una quantitat d'informació extraordinàriament gran que pot limitar de manera important la rapidesa en qualsevol tasca. A més, moltes de les cel·les contenen zeros, ja que hi ha moltes paraules amb les quals una paraula no coapareix. És per això que actualment hi ha gran interès en desenvolupar mètodes per **reduir dimensions** i així treballar amb un espai més manejable. La idea principal consisteix en crear un nou espai amb un nombre de dimensions menor fusionant o combinant aquelles files i columnes que siguin més semblants. Com a resultat, s'obté un espai que és més difícil d'interpretar directament, atès que les files ja no són paraules, sinó "classes latents". Aquesta reducció de dimensions proporciona, també, una reducció de soroll i de coordenades-zero, alhora que permet trobar coocurrències anomenades "d'ordre superior", és a dir, paraules similars que no coapareixen en contextos *idèntics*, però sí en contextos *similars* (Turney i Pantel, 2010; Landauer i Dumais, 1997). Aquesta aproximació es

coneix com a *Latent Semantic Analysis*. La reducció es duu a terme amb diferents models matemàtics i computacionals com ara la *Singular Value Decomposition*, les *Xarxes Neuronals*, inferència Bayesiana, o models log-lineals, com els de l'eina *word2vec*[9]. Perquè aquests models tinguin bons resultats, però, és imprescindible comptar amb una quantitat de dades molt gran.

3.3 Composició amb VSMs

Com s'acaba de veure, els VSMs ens permeten modelar el significat de paraules, que s'han considerat fins ara de manera aïllada. Un pas endavant en aquest mètode seria, doncs, modelar el significat de sintagmes o, fins i tot, d'oracions. És dificil establir límits clars entre els significats de paraules i de sintagmes o frases. Molt sovint, per exemple, es pot expressar el mateix significat amb una paraula o amb més d'una (*badallar* vs *fer un badall*). Per tant, si les paraules tenen una representació distribucional, també l'han de tenir els sintagmes (Baroni, 2013). Si es pot modelar el significat de sintagmes en VSMs, llavors es pot també obtenir similituds entre sintagmes, la qual cosa va inevitablement lligada a processos de **paràfrasi**. Seria possible, també, identificar certs tipus i aspectes de les inferències.

Actualment ja s'han proposat diversos mètodes per modelar la composició. Les propostes més rellevants fins ara són la de Mitchell i Lapata (2010), que proposen models d'addició i multiplicació de vectors, i la de Baroni i Zamparelli (2010), que distingeixen entre categories-vector i categories-funció, i modelen, per exemple, un sintagma Nom-Adjectiu com un vector (nom) al qual s'aplica una funció (adjectiu) per obtenir un altre vector, en el mateix espai dels noms.

Grau de composicionalitat

Una aplicació interessant que pot tenir la composició en VSMs, molt relacionada amb l'apartat 2 d'aquest treball, és la de mesurar el **grau de composicionalitat** de les MWEs. Comptar amb una mesura que determina el grau de composicionalitat d'una expressió permetria diferenciar, d'una llista de candidats a col·locacions obtinguda a partir de mitjans estadístics, aquells candidats que es corresponen amb la noció lingüística de MWE. Un altre avantatge que ens proporcionaria és el d'establir una classificació de diferents tipus de MWEs, com proposen Fazly i Stevenson (2014), que identifiquen el grau de composicionalitat amb la similitud entre una expressió i els seus constituents: si *menjar pomes* és semblant a *menjar* i també és semblant a *pomes*, llavors és una expressió composicional. Aquest és un tema en el qual s'està treballant actualment i ja han sorgit diverses propostes al respecte.

[9] https://code.google.com/p/word2vec/ .

Maldonado-Guerra i Emms (2011) proposen algunes maneres de mesurar la composicionalitat mitjançant la distància entre una col·locació i cada un dels seus elements per separat, o bé a partir de la similitud entre la col·locació i la resta de contextos del seu nucli. Aquest últim cas es tractaria de comparar, per exemple, *menjar pomes* amb tots els contextos en què apareix *menjar* seguit d'un objecte que no sigui *pomes*: *menjar peres, menjar plàtans...*

Kiela i Clark (2013) desenvolupen un mètode no supervisat amb l'objectiu de mesurar la composicionalitat basat en la similitud d'un sintagma amb **sintagmes veïns** (*neighbour phrases*). S'obté un sintagma veí quan se substitueix, en el sintagma que ens ocupa, un dels constituents per una paraula relacionada. Per exemple, els veïns de *menjar barrets* podrien ser *consumir barrets* o *menjar pantalons*. Aquests autors observen que, mitjançant la proposta composicional de Mitchell i Lapata (2010), quan un sintagma no és composicional, és molt pròxim (o similar) als seus veïns; mentre que un sintagma composicional -*menjar pomes*- presenta més distància amb els seus veïns. A partir d'aquesta observació creen una mesura de composicionalitat basada en la distància amb els veïns. Treballen amb combinacions verb+nom i aconsegueixen resultats semblants a altres mètodes més complexos.

Per una altra banda, Vecchi et al. (2011) proposen una manera de detectar anomalies o incompatibilitats semàntiques en parells de nom+adjectiu (NA). Conclouen que la **llargada del vector** NA és més curta en combinacions anòmales i que en una expressió anòmala el vector NA es trobarà més **allunyat del nom N**.

L'objectiu d'aquests autors és determinar la inacceptabilitat d'una expressió no vista en el corpus de manera automàtica. Tot i així, les intuïcions que fan servir, aplicades a combinacions ja vistes, també podrien donar compte de la composicionalitat de les expressions, essent aquelles combinacions d'elements incompatibles més susceptibles de ser considerades com a MWEs en el sentit lingüístic.

4 VSMs per detectar MWEs: Un experiment

Fins ara hem vist dues aproximacions a l'anàlisi quantitativa del significat que apliquen diferents mètodes i tenen diferents objectius. Per una banda, l'extracció de MWEs amb mitjans estadístics i, per l'altra, l'ús de VSMs per representar el significat de paraules, sintagmes o frases. Martí et al. (2015) exploren l'ús de representacions basades en VSMSs per identificar MWEs. La metodologia que segueixen és el següent.

El corpus Diana-Araknion és un corpus de l'espanyol de 100 milions de paraules analitzat sintàcticament amb dependències. A partir dels 10.000 lemes[10] més freqüents del corpus, es crea una matriu paraula-context en la qual el context consisteix en dependències sintàctiques representades per una relació, la direcció de la relació (pare-fill) i la paraula relacionada. Per exemple:

El barbero afeita la larga barba de Jaime
un context de *barba* és c_1 = [<:dobj:afeitar]

Els autors parteixen de la **hipòtesi patró-construcció**: aquells contextos que siguin rellevants per definir un conjunt (clúster) de paraules semànticament relacionades tendeixen a ser parts de construccions lexicosintàctiques. El següent pas consisteix en la clusterització (agrupament) dels lemes en base als contextos (trets) que comparteixen. Com a resultat, es van obtenir 700 clústers de paraules semànticament relacionades. En el 97% dels clústers, les paraules que els componen pertanyen a la mateixa categoria. Posteriorment, s'estableixen relacions entre clústers a partir de la informació continguda en els contextos, donant lloc a un graf de clústers relacionats. Finalment, es generen candidats a patró combinant cada una de les paraules dels clústers amb cada una de les paraules dels clústers relacionats.

L'objectiu d'establir relacions entre clústers és el de trobar parells de paraules que estiguin relacionades tot i no coaparèixer de manera directa en el corpus.

De tots els clústers relacionats, apliquen un filtre per quedar-se només amb aquells que mantenen entre ells una relació bidireccional o commutativa, per garantir la qualitat dels patrons resultants. Com a resultat d'aquest procés, els candidats a patrons lexicosintàctics que s'obtenen són del tipus:

barba <:dobj afeitar
accidentar >:subj aeronave
pesadilla >:atr desagradable

La qualitat d'aquests candidats a patró és avaluada a partir de judicis humans, preguntant si es corresponen amb bons patrons lingüístics. S'ha obtingut un 88,3% de respostes positives.

[10] Es limiten als noms, adjectius, verbs, adverbis i preposicions.

Tot i els bons resultats, es va observar que alguns dels clústers es relacionaven incorrectament a causa de paraules polisèmiques. Un exemple és el del clúster 24, que es relaciona amb el clúster 276, tot i no guardar cap mena de relació semàntica:

clúster 24: *barba, bigote, cabellera, cabello, cana, ceja, hebra, mecha, mechón, melena, pelo, peluca, pestaña, rizo, trenza.*

clúster 276: *arbusto, castaño, encina, follaje, haya, laurel, maleza, mata, matorral, palmera, pino, rama, roble, sauce, tronco, álamo, árbol.*

Aquests dos clústers es van relacionar l'un amb l'altre per la paraula polisèmica *castaño*, que és un context comú dels dos clústers, perquè pot referir-se tant al color del cabell (clúster 24) com a l'arbre (clúster 276).

Així doncs, si s'aconseguís solucionar la representació de la polisèmia en aquest model, s'aconseguirien millors clústers i millors relacions entre clústers.

5 Polisèmia en VSMs

En els VSMs, a cada paraula li correspon un únic vector. Això significa que en el cas de les paraules polisèmiques, es poden tenir múltiples sentits codificats en un sol vector. Això pot portar problemes, ja que no queda representada una propietat del significat lèxic: la polisèmia.

En aquest apartat presento una proposta per tractar la polisèmia en el marc dels *Vector Space Models*. El nostre objectiu és induir sentits de manera automàtica a partir del text, per aconseguir finalment vectors de **sentits** i no de **paraules**.

Fins ara ja hi ha hagut propostes per solucionar el tractament de la polisèmia però són, ara per ara, insuficients.

Una via de solució és la **clusterització de contextos** (Schütze, 1998). La proposta consisteix a prendre tots els tokens -o ocurrències- de la paraula objectiu directament del corpus i es crea un vector per a cada ocurrència d'aquesta paraula amb el seu corresponent context, obtenint un vector per cada token de la paraula *target* al corpus. Posteriorment, clusteritza aquests vectors segons la seva similitud, assumint que cada clúster és un sentit de la paraula.

Pantel i Lin (2002), en canvi, proposen un **clústering de paraules**. Creen, per a cada paraula *target*, un clúster amb les paraules més similars -o més properes en l'espai- a partir d'un VSM. A continuació, en un procés recursiu, s'assigna a cada paraula el clúster que més se li assembla, que correspondrà a un sentit. Llavors s'elimina del clúster de la paraula target aquells trets que tenen en comú amb el clúster que se li acaba d'assignar, i es repeteix el pas anterior per trobar el segon clúster més similar.

5.1 Proposta de tractament de la polisèmia en VSMs

Com a aportació personal en aquest treball, proposo una metodologia per a la representació de la polisèmia en els VSMs. La proposta es basa en la següent hipòtesi:

"*Els contextos d'un sentit d'una paraula són similars entre ells*"

Cal remarcar la diferència amb Schütze (1998), que treballa amb **contextos de tokens**, mentre que els contextos que nosaltres tractem són **contextos de lemes** o, més concretament, contextos de types de lemes.

En primer lloc, el mètode que proposo requereix d'una representació en el format dels VSMs d'un corpus amb una matriu del tipus paraula-context en la qual el context siguin lemes que han coaparegut amb els altres. Explicarem el procediment amb un exemple inventat de poques dimensions. Suposem que tenim

el vector de la paraula ambígua *cor*. Aquesta podria ser una llista dels seus contextos més freqüents:

COR: (òrgan, passió, fetge, pedra, interpretar, infart, director, amor, ànima, malalt)

Veiem que COR té com a contextos paraules molt diferents. Això, en l'Espai Semàntic, es reflectirà en un vector sense una direcció clara, i probablement de poca llargada. Podem fer un reordenament d'aquests contextos:

COR: (òrgan, fetge, infart, malalt, amor, ànima, passió, pedra, director, interpretar)

Ara ja intuïm de manera directa els diferents sentits de cor: el cor com a òrgan, com a "gestionador" de sentiments, com a grup de cant. I *pedra*? *Pedra* podria venir de l'expressió metafòrica *tenir un cor de pedra*.

L'objectiu és partir el vector difús de COR en tres vectors diferents, amb una direcció més definida: COR1, COR2, COR3. Segons la hipòtesi de partença, si els contextos dels diferents sentits d'una paraula són similars, llavors es pot clusteritzar per similitud tots els contextos de *cor*. *Òrgan, fetge, amor...* tots ells són paraules que també han aparegut alguna vegada al corpus, per tant tenen el seu propi vector, i podem calcular la similitud entre vectors amb el seu cosinus. És probable que *òrgan* i *fetge* s'assemblin entre ells, i que *malalt* s'assembli a *infart*, etc. Així, agrupant-los, podria obtenir tants clústers com sentits té la paraula i partir el vector en les parts corresponents. Llavors, si abans teníem una matriu com la de la Taula 2[11], ara en tenim una com la de la Taula 3. Cal decidir què fer si hi ha algun context que no s'assembla especialment a res (*pedra*). Podem decidir, per defecte, mantenir un context com *pedra* en tots els sentits que es formin.

	òrgan	fetge	infart	malalt	amor	ànima	passió	pedra	director	interpretar	...
cor	x	x	x	x	x	x	x	x	x	x	...
òrgan	x	x	x	x	x	x	x	x	x	x	...
fetge	x	x	x	x	x	x	x	x	x	x	...
...	...	...	...	...	...	...	...	...	...	...	

Taula 2: Matriu original

Així doncs, passem a tenir vectors **de sentits**, i no simplement **de paraules**.

[11] Tant a la Taula 2 com a la Taula 3, x és la puntuació de coaparició donada per una mesura d'associació i representa qualsevol valor estrictament més gran que 0. Entenem per 0 la nul·la coaparició i suposem que la mesura només dóna valors positius o 0.

	òrgan	fetge	infart	malalt	amor	ànima	passió	pedra	director	interpretar	...
cor1	x	x	x	x	0	0	0	x	0	0	...
cor2	0	0	0	0	x	x	x	x	0	0	...
cor3	0	0	0	0	0	0	0	x	0	0	...
...	...	...	...	...	...	...	...	...	...	...	

Taula 3: Matriu de sentits

Idealment, això s'hauria de dur a terme amb totes les paraules de la matriu. Un inconvenient que pot tenir aquest mètode és que no permet desambiguar els contextos. Això significa que les paraules que tenen *cor* com a context s'hauran de desambiguar sense poder saber amb quin sentit de *cor* van coaparèixer (amb COR1, COR2, o COR3), ja que els sentits no es descobreixen sobre instàncies de la paraula en el corpus (tokens), sinó ja sobre la matriu.

No es tracta d'un procés trivial, per molt senzilla que sembli la intuïció que hi ha al darrere. Vegem-ne els passos:

1. El primer pas consisteix a crear una **matriu de similituds**. Aquesta matriu la obtindrem calculant la similitud de tots els possibles parells de paraules. Serà una matriu quadrada i simètrica. L'objectiu de construir aquesta matriu des d'un bon començament és estalviar-se successius i repetitius càlculs posteriors. D'aquesta manera només cal calcular cada similitud una vegada.
   A continuació cal processar paraula per paraula per obtenir-ne els sentits. Seguirem l'explicació amb l'exemple de *cor*.

2. Prenem tots els contextos de cor que superin un **llindar de freqüència** determinat. Suposarem que aquests són els 10 contextos presentats en l'exemple de *cor*. Per ara només volem identificar sentits generals i per això podem prescindir dels contextos de poca freqüència.

3. A la matriu de similituds, extraiem tots els valors que siguin del nostre interès, és a dir, la similitud de tots els contextos de la paraula entre ells: *òrgan* amb *fetge*, *òrgan* amb *infart*, *òrgan* amb *malalt*... Com que tenim 10 contextos rellevants, si ignorem la similitud d'un element amb ell mateix, tenim 45 parells, cada un amb el seu valor de similitud.

4. El següent pas consisteix en la clusterització dels elements més similars. Aquest procediment es pot dur a terme amb un algoritme d'**agrupament jeràrquic**:

   (a) Establim un **llindar de similitud** que cal superar per poder pertànyer al mateix clúster. En primer lloc s'ajunten els dos elements que tinguin la similitud més alta de totes. En aquest punt, passem de tenir 10 **elements** a

tenir 9 **grups**, candidats a clúster, només un dels quals està format per dos elements (o paraules). Per exemple, pot ser el grup *director, interpretar*.

(b) Com que només tenim mesures de similitud entre paraules, ens cal establir una **mesura de similitud entre grups de paraules**. Algunes opcions típiques són prendre la **similitud màxima** entre els elements del grup, la **mínima** o la **mitjana**. En el nostre exemple, en el cas de la similitud màxima, prendrem com a similitud entre el grup *director, interpretar* i *òrgan*, la similitud que hi ha entre l'element *òrgan* i l'element del grup que sigui més semblant a òrgan, per exemple, *director*.

(c) Un cop hem definit aquesta nova mesura de similitud, el procediment consistirà en ajuntar a aquest primer clúster els elements que més se li assemblin, un per un. Quan ja no quedin elements que superin el llindar de similitud amb el clúster, es donarà el clúster per acabat. Com que segons el llindar de similitud cap altra paraula més s'assembla a *director, interpretar*, ja s'ha format un sentit o un clúster definitiu.

(d) Es repeteix el procés amb els següents dos elements més semblants entre ells que estiguin fora del clúster ja creat, creant-ne un de nou. Per exemple, s'ajunten *fetge* i *malalt*, formant el grup *fetge, malalt*. S'apliquen els passos (c) i (d) diverses vegades.

(e) El procés s'atura en tres moments possibles: si ja no queden grups o elements que superin el llindar de similitud, o si s'ha arribat ja a un nombre màxim de sentits que podem determinar nosaltres segons la granularitat que vulguem obtenir.

(f) Finalment, cal decidir què es fa amb aquells elements que un cop acabat el procés no han estat assignats a cap sentit, per ser molt diferents a la resta (*pedra*). Una proposta és incloure'ls a tots els sentits. En aquest cas caldria disminuir la seva puntuació en la matriu original, dividint-la per exemple pel nombre de sentits. Altrament, estaríem alterant excessivament els recomptes de freqüència de la matriu. Una altra alternativa seria que cada un formés un clúster d'un sol element, com si tingués el seu propi sentit, però no seria la millor opció en cas que n'hi haguessin molts.

5. A partir d'aquest agrupament, podem modificar la matriu i crear tants vectors de *cor* com l'algoritme ens hagi proposat, tal i com s'ha mostrat abans: COR1, COR2 i COR3.

Resum de característiques

Aquest és un resum de les característiques principals d'aquesta proposta i una idea de com es diferencia d'allò que s'ha fet fins ara:

- A diferència dels sistemes habituals de *Word Sense Disambiguation*, o desambiguació de sentits de paraules, no pretén desambiguar paraules en el text, sinó **descobrir sentits de les paraules a partir del seu ús**. Per aconseguir-ho no fa servir recursos fets a mà, de manera que no passa per cap filtre de percepció subjectiva, sinó que indueix possibles sentits a partir de les dades de la matriu. És, per tant, una proposta de *Word Sense Induction* o d'inducció de sentits de paraules.

- A diferència de la proposta de Pantel i Lin (2002), no treballa amb clústers prèviament definits que puguin limitar d'entrada el nombre de sentits a descobrir. El nombre de sentits és, des de l'inici, obert.

- A diferència de la proposta de Schütze (1998), no treballa amb tokens -instàncies d'una paraula al corpus-, sinó amb types de lemes directament trets de la matriu de coocurrència.

- L'única restricció en el nombre de sentits que pot tenir una paraula és l'establiment un nombre màxim. Cada paraula té el seu nombre de sentits, no necessàriament igual al de les altres, indeterminat a l'inici i variable segons els paràmetres.

- No és imprescindible que tots els contextos estiguin en un únic sentit: poden quedar contextos inclassificables (*pedra*).

5.2 Implementació de la proposta

Ha estat possible implementar la proposta del tractament de la polisèmia sobre les matrius del Diana-Araknion amb algunes modificacions respecte de com estava pensada. Aquesta prova està basada en la mateixa intuïció que la proposta però té diferències pel que fa a l'execució:

- S'ha establert un nombre predefinit de sentits per paraula. Provant amb 3 clústers, l'algoritme de clusterització separava les paraules per categoria gramatical. Per tal d'obenir dades més rellevants, s'estableix l'obtenció de 21 clústers per paraula.

- No s'ha dut a terme un agrupament jeràrquic.

- No s'ha tingut en compte directament la mesura de similitud, sinó la coincidència de contextos.

S'ha dut a terme amb 6 paraules del corpus, algunes clarament polisèmiques (*banco*) i d'altres, no (*médico*). Observant els resultats, resulta evident que 21 és un nombre de sentits excessiu. Si bé en la majoria de casos les paraules d'un únic clúster pertanyen clarament a un únic sentit, hi ha diversos clústers que es refereixen a un mateix sentit. És el cas de *banco*. En el següent exemple

veiem dos clústers clarament pertanyents a un sentit (institució econòmica), que desitjablement n'haurien de formar un de sol; i un de tercer, pertanyent a un altre (seient):

Clúster 5: *agencia, banco, comercio, compañía, empresa, entidad, filial, industria, institución, internet, negocio, oficina, operador, sector, seguro, servicio, tecnología, telecomunicación.*

Clúster 17: *accionista, activo, asalto, cajero, capitalización, cliente, depósito, directorio, empleado, gobernador, privatización, quiebra, reestructuración, rescate, robo, saneamiento, solidez, sucursal.*

Clúster 20: *caja, casa, cocina, estación, exterior, extremo, fila, fondo, jardín, lado, mesa, nave, parque, paseo, pie, plaza, popa, puerta, punta, segundo, silla, tienda.*

A la llum dels resultats es poden treure les següents conclusions:

- El fet que la separació en 3 només separés paraules per categoria gramatical pot estar causat pel fet d'haver utilitzat **coaparició sintàctica**. Les paraules d'una categoria s'assemblen més entre elles perquè tendeixen a tenir el mateix tipus de relacions sintàctiques. És possible que sigui millor tenir en compte la coaparició **superficial** o **textual**.

- Alguns clústers estan formats per paraules que no són discriminadores de sentits. És el cas del següent clúster de la paraula *gato*:

  Clúster 14: *aparecer, desaparecer, encontrar, entrar, pasar.*

  És possible que sigui necessari **desestimar paraules que tinguin alta similitud amb la major part de les altres** paraules, pel fet que no ajudarien a distingir un sentit d'un altre.

- És necessari **no forçar un nombre de clústers predefinit**, especialment pels casos de paraules poc o gens polisèmiques com *médico* o *perro*.

La intuïció en què es basa la proposta sembla anar per bon camí, ja que en els clústers formats, en general, tenen molta **cohesió interna**.

6 Limitacions i feina futura

A continuació presento breument diferents línies obertes de recerca en el marc de la detecció de MWEs i de la representació del significat amb VSMs.

Detecció estadística de MWEs

La hipòtesi nul·la d'aleatorietat de les dades lingüístiques es troba excessivament allunyada de la realitat, ja que les paraules mai es troben ordenades aleatòriament. Això pot donar lloc a puntuacions d'associació inflades, és a dir, pot fer que una coaparició casual tingui més probabilitat de ser considerada significativa. Una possible solució seria trobar una nova hipòtesi nul·la que tingui en compte algunes restriccions combinatòries de la llengua (Evert, 2008). Aquesta proposta es veu reforçada pel fet que les coaparicions entre dues paraules són encara menys freqüents que les paraules en si.

En els models presentats, la cerca se centra en col·locacions, que són combinacions de dues paraules. Sovint, les col·locacions que es troben són, de fet, fragments de MWEs que contenen més paraules. Caldria desenvolupar un mètode que ens permetés anar més enllà de les combinacions de dues paraules i trobar-ne de més extenses. L'ideal seria trobar un sistema que detectés automàticament la independència o el solapament de les combinacions de paraules (Evert, 2008). Això ens permetria, també, descobrir patrons formals (no plens substantivament) que no serien possibles d'identificar estadísticament amb els mitjans actuals.

Un altre tema que queda pendent és el de les **col·locacions assimètriques**, que són aquelles en les quals un dels elements és altament predictible a partir de l'altre, però no a l'inrevés. Per exemple, en *la Ilíada*, si coneixem el segon element (*Ilíada*), el primer (*el*) és completament predictible, però no passa el mateix a l'inrevés. S'han proposat mètodes basats en probabilitats condicionals, però encara no obtenen resultats prou bons (Michelbacher et al., 2007).

VSMs

Els VSMs constitueixen actualment una àrea de recerca que rep molta atenció i es troba en ràpid i constant progrés. Hi ha encara molts aspectes per millorar:

- Cal encara una aproximació que sigui capaç de tractar la **polisèmia**. Caldria veure els resultats de la proposta feta en aquest treball després de les modificacions especificades,
- Cal treballar més en un model de la **composició** òptim,
- Cal trobar un mètode fiable per calcular el **grau de composicionalitat** d'una MWE, tasca que correspon als VSMs però seria de gran ajut a la detecció de MWEs.

7 Conclusió

En aquest treball s'han presentat dues noves aproximacions al significat per mitjans computacionals basades en estadística i àlgebra: la detecció de MWEs i l'ús de VSMs per a la representació del significat. S'ha presentat un experiment on es combinen l'objectiu de la primera amb el mètode de la segona, per trobar patrons del llenguatge utilitzant VSMs. Finalment, s'ha fet una proposta de tractament de la polisèmia en el marc dels VSMs que podria millorar el resultat d'aquest experiment i d'altres tasques basades en VSMs.

Referències

Baldwin, Timothy, Emily M. Bender, Dan Flickinger, Ara Kim, i Stephan Oepen (2004), "Road-testing the english resource grammar over the british national corpus." A *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 2047–2050.

Baroni, Marco (2013), "Composition in distributional semantics." *Language and Linguistics Compass*, 7, 511–522.

Baroni, Marco i Alessandro Lenci (2010), "Distributional memory: A general framework for corpus-based semantics." *Comput. Linguist.*, 36, 673–721.

Baroni, Marco i Roberto Zamparelli (2010), "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space." A *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, 1183–1193, Association for Computational Linguistics, Stroudsburg, PA, USA, URL `http://dl.acm.org/citation.cfm?id=1870658.1870773`.

Boleda, Gemma i Kathrin Erk (2015), "Distributional semantic features as semantic primitives –or not." *AAAI Spring Symposium on KRR (Knowledge Representation and Reasoning)*.

Bruni, E., Tran K., i Baroni M. (2014), "Multimodal distributional semantics." *Journal of Artificial Intelligence Research*, 1–47.

Croft, W. i D.A. Cruse (2004), *Cognitive Linguistics*. Cambridge Textbooks in Linguistics, Cambridge University Press, URL `http://books.google.es/books?id=I6Z9H-eRSgoC`.

Curran, James i Marc Moens (2002), "Scaling context space." A *Proceedings of ACL*, 231–238.

Dunning, Ted (1993), "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics - Special issue on using large corpora*, 19, 61–74.

Evert, Stefan (2008), *Corpus Linguistics. An International Handbook*, chapter Corpora and collocations, 1212–1248. Mouton de Gruyter.

Fazly, Afsaneh i Suzanne Stevenson (2014), "A distributional account of the semantics of multiword expressions." *Italian Journal of Linguistics*, 158–179.

Fillmore, Charles J., Paul Kay, i Mary Catherine O'Connor (1988), "Regularity and idiomacity in grammatical constructions: The case of let alone." *Language*, 64, 501–538.

Firth, John R. (1957), *Papers in Linguistics, 1934-1951*. Oxford University Press, Oxford, UK.

Fodor, J., J. A. Fodor, i M. F. Garret (1975), "The psychological unreality of semantic representations." *Linguistic Inquiry*, 6, 515–531.

Harnad, Stevan (1990), "The symbol grounding problem." *Physica D: Nonlinear Phenomena*, 146–162.

Harris, Zellig (1954), "Distributional structure." *Word*, 10, 146–162.

Katz, J.J. i J.A. Fodor (1963), "The structure of a semantic theory." *Language*, 39, 170–210.

Kiela, D. i S. Clark (2013), "Detecting compositionality of multi-word expressions using nearest neighbours in vector space models." A *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-13)*.

Landauer, Thomas K i Susan T Dumais (1997), "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, 104, 211.

Lenci, Alessandro (2008), "Distributional semantics in linguistic and cognitive research." *Italian Journal of Linguistics*, 1–31.

Maldonado-Guerra, Alfredo i Martin Emms (2011), "Measuring the compositionality of collocations via word co-occurrence vectors: Shared task system description." A *Proceedings of the Distributional Semantics and Compositionality workshop (DISCo 2011)*.

Martí, M. Antònia, Manuel Bertran, Mariona Taulé, i Maria Salamó (2015), "Distributional approach based on syntactic dependencies for discovering constructions." *Computational Linguistics*. En procés d'evaluació.

Michelbacher, Lukas, Stefan Evert, i Hinrich SchÃijtze (2007), "Asymmetric association measures." *RANLP*.

Miller, George i Walter Charles (1991), "Contextual correlates of semantic similarity." *Language and Cognitive Processes*, 6.

Mitchell, Jeff i Mirella Lapata (2010), "Composition in distributional models of semantics." *Cognitive Science*, 34, 1388–1439.

Moore, Robert C. (2004), "On log-likelihood-ratios and the significance of rare events." A *Proceedings of EMNLP 2004* (Dekang Lin i Dekai Wu, eds.), 333–340, Association for Computational Linguistics.

Nunberg, Geoffrey, Ivan A Sag, i Thomas Wasow (1994), "Idioms." *Language*, 491–538.

Pantel, Patrick i Dekang Lin (2002), "Discovering word senses from text." A *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*, 613–619.

Pecina, Pavel (2010), "Lexical association measures and collocation extraction." *Language Resources and Evaluation*, 44, 137–158.

Pulvermüller, F. (2005), "Brain mechanisms linking language and action." *Nature Reviews Neuroscience*, 576–582.

Salton, G., A. Wong, i C.S. Yang (1975), "A vector space model for automatic indexing." *Communications of the ACM*, 18, 613– 620.

Schütze, H. (1998), "Automatic word sense discrimination." *International Journal of Corpus Linguistics*, 24, 97 – 123.

Turney, Peter D. i Patrick Pantel (2010), "From frequency to meaning: Vector space models of semantics." *J. Artif. Int. Res.*, 37, 141–188, URL `http://dl.acm.org/citation.cfm?id=1861751.1861756`.

Vecchi, Eva Maria, Marco Baroni, i Roberto Zamparelli (2011), "(linear) maps of the impossible: Capturing semantic anomalies in distributional space." A *Proceedings of the ACL Workshop on Distributional Semantics and Compositionality*, 1–9.

Wehrli, Eric (2014), "The relevance of collocations for parsing." A *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, 26–32, Association for Computational Linguistics.

Wible, David i Nai-Lung Tsao (2010), "Stringnet as a computational resource for discovering and investigating linguistic constructions." A *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, EUCCL '10, 25–31, Association for Computational Linguistics, Stroudsburg, PA, USA, URL `http://dl.acm.org/citation.cfm?id=1866732.1866736`.

Wittgenstein, Ludwig (1953), *Philosophical Investigations*. Blackwell, Oxford, UK.

Wray, Alison i Mick Perkins (2000), "The functions of formulaic language: an integrated model." *Language and Communication*, 20, 1–28.