Judit Feliu i Mireia Trias (cur.)
Viquipèdia i terminologia
Barcelona: Institut d'Estudis Catalans, 2021, p. 15-47 DOI: 10.2436/15.2503.02.64

Viquipèdia: un recurs útil per a la terminologia?

Jorge Vivaldi Palatresi
Institut de Lingüística Aplicada (Universitat Pompeu Fabra)

1. Introducció

El projecte Viquipèdia (WP, de l'anglès Wikipedia) es pot considerar un dels més reeixits pel que fa a la recopilació de coneixement. Des dels seus inicis, l'any 2001,[1] fins ara, el nombre de lectors ha anat augmentant vertiginosament. Actualment, està disponible en més de tres-centes llengües[2] i el nombre de lectures de pàgines en tot el món es compten per milions cada dia.[3] Aquesta fita ha sigut possible gràcies a la gestió de la Fundació Wikimedia i a milers de persones que hi han contribuït i contribueixen amb el seu coneixement i el seu temps.

L'aproximació que segueix aquest desenvolupament, a diferència dels tradicionals, és que es tracta d'un projecte obert. Això significa que qualsevol persona que disposi d'un ordinador i de connexió a Internet pot fer-hi una contribució, ja sigui escrivint un article o modificant-ne un de ja existent. La filosofia de la Viquipèdia és que si una comunitat treballa conjuntament en el contingut d'un article o d'un àmbit, aquest millorarà amb el temps. D'aquesta manera es podria dir que un article mai no està acabat, ja que potencialment pot ésser modificat en qualsevol moment. En general, el projecte presenta una resposta molt ràpida (i superior a la de qualsevol altre recurs similar) als esdeveniments més rellevants. Aquest mode de treball fa possible que hi hagi articles vandàlics, és a dir, amb informació dubtosa o bé que responen a altres interessos. Viquipèdia ha anat modi-

1. En la pàgina «Viquipèdia en català» es poden consultar moltes dades sobre els orígens d'aquest recurs.
2. Vegeu estadístiques actualitzades a *https://en.wikipedia.org/wiki/List_of_Wikipedias*.
3. Vegeu les estadístiques globals a *https://stats.wikimedia.org/EN/*.

ficant la política d'admissió d'articles i de revisions per tal de protegir-se d'aquests problemes.

L'èxit aconseguit ha atret recercadors de molts àmbits. En conseqüència, existeix una literatura abundant en què s'analitzen diversos aspectes d'aquest projecte des de punts de vista diferents. Cal mencionar, entre altres, els treballs de revisió de la literatura existent de [1] i [2]. També existeixen dues plataformes que recullen treballs de diferents tipus: WikiLit[4] i WikiPapers.[5]

L'impacte de la Viquipèdia abasta diferents aspectes del saber que inclouen la representació del coneixement, la sociologia o l'educació, entre altres. Aquesta presentació se centrarà en aspectes relatius a l'explotació de la Viquipèdia com a font de coneixement per a projectes de processament del llenguatge natural (PLN).

Després d'aquesta introducció, en l'apartat 2, estudiarem amb cert detall l'estructura interna de la Viquipèdia. A continuació, en l'apartat 3, analitzarem breument una qüestió complexa i controvertida com és la credibilitat. En l'apartat 4, mostrarem com es poden fer consultes sistemàtiques en aquest recurs i continuarem amb l'apartat 5, en què analitzarem algunes qüestions a tenir en compte quan es fan consultes informàtiques a la Viquipèdia. En l'apartat 6, presentarem molt breument algunes aplicacions d'aquest recurs en diferents aspectes del PLN, entre els quals s'analitzaran amb cert detall algunes aplicacions de l'àmbit de la terminologia. Finalment, presentarem algunes conclusions que es poden extreure d'aquest treball.

2. Estructura de la Viquipèdia

Un usuari qualsevol pot consultar molt fàcilment qualsevol article de la Viquipèdia. Una observació atenta de qualsevol pàgina revelarà l'existència d'una gran quantitat d'informació associada. Aquest fet hauria de fer pensar que aquest recurs disposa d'una estructura capaç i prou complexa per respondre a la majoria de les necessitats d'informació de qualsevol usuari. En aquest apartat es presenta amb cert detall l'estructura de dades associada a aquest recurs.

En primer lloc, cal assenyalar que la unitat d'informació de la Viquipèdia és l'article. L'usuari, quan fa una consulta, a través d'una pàgina web, rep una pàgina com a resposta. Diferenciem *article* de *pàgina* en el sentit que l'article conté estrictament un text amb l'explicació enciclopèdica d'una unitat d'informació. La pàgina, en canvi, conté, a més de l'article, altra informació relativa a aquesta unitat, com ara les categories a les quals està vinculada, la traducció a altres llengües, bibliografia rellevant, pàgines web on es pot ampliar la informació, etc.

4. *http://wikilit.referata.com/wiki/Main_Page*.
5. *http://wikipapers.referata.com/wiki/Main_Page*.

En el text de cada article hi ha alguna paraula (o grup de paraules) que serveixen també com enllaços a altres articles de la mateixa llengua. En cada article hi ha, de mitjana, quinze enllaços d'aquest tipus. L'autor de l'article, d'acord amb la guia d'estil, escull quines són la paraula o les paraules que considera necessari associar a un enllaç per tal de facilitar la comprensió de l'article. Es tracta, doncs, d'una informació que, en potència, és semànticament rellevant. Aquests enllaços són unidireccionals i es denominen *enllaços de sortida*. De la mateixa manera, cada pàgina és apuntada per un cert nombre de paraules d'altres pàgines, el conjunt de les quals se solen denominar *enllaços d'entrada*. El conjunt d'articles i els seus enllaços formen un graf dirigit.[6]

Cada article té assignades una o més categories mitjançant el que s'acostuma a denominar *enllaços categorials*. Aquestes categories es poden veure com a classes que tenen associades una sèrie d'instàncies, que en aquest cas són pàgines. Al mateix temps, una categoria està vinculada a altres categories mitjançant enllaços anomenats *supercategories* (categories que trobem quan recorrem el graf cap al *top*) o *subcategories* (resta de categories). Encara que no ho siguin sempre, és freqüent considerar aquests enllaços com a taxonòmics.

Podem veure el conjunt de les categories com a un altre graf dirigit (en [3] trobareu una interessant anàlisi dels dos grafs). Els nusos d'aquest graf són les categories associades a cada pàgina i els enllaços són els vincles entre aquestes categories. Segons [4] les categories es poden classificar de la manera següent:

1. Categories de contingut (*content categories*): categories destinades a ajudar l'usuari a trobar articles segons atributs d'aquests. Es poden dividir en aquests subgrups:

— Categories de temes (*topic categories*). Per exemple, la «Categoria: Catalunya» conté els articles relacionats amb el tema *Catalunya*.

— Categories de conjunts (*set categories*): categories que indiquen una classe, normalment en plural. Per exemple, la «Categoria: Automòbils» conté els articles relacionats amb el tema *automòbils*.

2. Categories de projecte o de servei (*project categories*): categories destinades a l'organització interna del projecte i que són utilitzades per editors o per eines automàtiques. En són exemples les categories ocultes, els esborranys d'articles, els articles que necessiten neteja o ampliació, etc.

A més dels enllaços ja mencionats, una pàgina de la Viquipèdia pot contenir altres tipus d'informació i enllaços, com ara:

6. En general, es pot dir que un *graf* és una representació abstracta d'un conjunt d'objectes (o nodes); alguns parells d'aquests objectes estan connectats per arestes. En aquest cas, els objectes són les pàgines i les arestes són enllaços que connecten paraules amb pàgines. Es diu que un graf és dirigit quan les arestes que uneixen els nodes tenen una orientació definida.

— URL externs: adreces web que són potencialment interessants en relació amb l'article que s'està visualitzant (p. ex., «grip» → «canal Salut específic de la Generalitat de Catalunya»).

— InterViqui: article, en una altra llengua, que presumiblement és equivalent a l'article que s'està visualitzant.[7]

— Altres articles de la Viquipèdia que amplien la informació amb temes fortament relacionats amb la pàgina que s'està visualitzant (p. ex., «Galileo Galilei» → «termòmetre de Galileu»).

— Bibliografia de referència. Pot estar indicada en forma de referència bibliogràfica o bé mitjançant un URL.

— Informació potencialment rellevant, sovint fa referència a URL.

— Registre d'autoritat. Indicació de les fonts d'informació utilitzades per a la confecció de l'article.

La figura 1 mostra esquemàticament l'estructura global de la Viquipèdia. La part dreta mostra esquemàticament una pàgina qualsevol i les relacions més rellevants que té associades, mentre que la part esquerra reflecteix la posició de cadascuna de les categories associades a aquesta pàgina i com aquestes es relacionen amb altres categories del graf de categories.

Totes les categories del graf de categories tenen associades almenys una pàgina. En la figura 1, l'article de la pàgina que es detalla en el graf de la dreta té associades altres pàgines («pàgina a», «pàgina b»…) cadascuna de les quals té una estructura semblant a la pàgina que es mostra. És a dir, cadascuna d'elles té associades categories i es relacionen amb altres pàgines. El mateix succeeix amb la resta d'informacions.

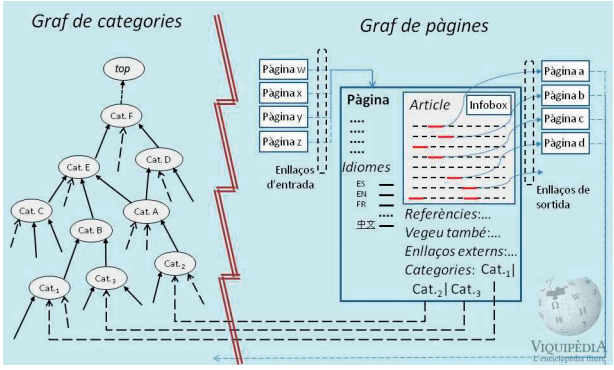

FIGURA 1. Estructura de pàgines i categories de la Viquipèdia.
FONT: Elaboració pròpia.

7. Cal comentar que molts articles són una traducció adaptada i/o retallada de l'article equivalent en llengua anglesa.

A la Viquipèdia existeixen també alguns altres tipus d'informacions de gran utilitat però que no es fan evidents a primer cop d'ull:

1. Pàgines de redirecció (*redirect pages*): informacions emmagatzemades a l'estructura de la Viquipèdia que permeten resoldre casos de:

— Sinonímia. Es redirigeix cap a la pàgina principal. Per exemple: «febres tercianes» → «malària», o «bicicle» → «velocípede».

— Equivalències. Mots que es consideren equivalents. Per exemple: «ronyons» o «sistema renal» → «ronyó».

— Desplegament de sigles no ambigües. Per exemple: «SCATERM → Societat Catalana de Terminologia».

— Correspondència entre un adjectiu relacional i un nom. Per exemple: «hepàtic» → «fetge», o «apical» → «àpex»).

— Correcció d'alguns errors tipogràfics més comuns que pot cometre l'usuari. Un exemple d'aquesta situació és quan l'usuari interroga aquest recurs amb les paraules «ronyo» o «ronyò» i el sistema mostra directament la pàgina amb la grafia correcta de «ronyó».

2. Pàgines de desambiguació (*disambiguation pages*). Es tracta de casos d'homonímia; és a dir, de pàgines amb títols molt semblants o semànticament ambigües. L'exemple clàssic d'aquesta situació és el mot «banc», que en la Viquipèdia incorpora sis lectures. Un altre exemple és l'entrada de «cosa», que inclou, entre altres, una referència a l'objecte, però també a l'organització Cosa Nostra o al personatge Juan de la Cosa. Aquestes pàgines s'utilitzen també per al desplegament de sigles ambigües com ara «IPC», per a la qual s'indiquen els tres significats possibles.

Finalment, si el que se cerca és informació terminològica hi ha una peça d'informació que és molt rellevant: la *infobox* (o infotaula). Aquesta informació normalment es mostra en forma d'una taula que apareix en l'angle superior dret de l'article i serveix per mostrar-ne (en forma de parells atribut-valor) un resum dels aspectes més rellevants. Sovint inclou fotografies, esquemes, etc. La informació acumulada a les *infoboxes* són una peça fonamental d'informació que són captades per la DBpedia i altres usuaris especialitzats.

Una qualitat de les *infoboxes* és que les dades que s'hi inclouen faciliten la comparació entre articles semblants. Per exemple, en medicina totes les malalties tenen informacions comunes, com ara: especialitat que la tracta, símptomes, medicació, part del cos afectada, causa, efectes, codis de classificació (CIM, CIAP), recursos externs que l'analitzen (MeSH, MedicinePlus, SNOMED-CT), etc. D'aquesta manera és fàcil i ràpid trobar informació pròpia d'una malaltia i comparar-la amb la d'altres. Cada branca de la ciència té definides les seves peces d'informació específiques. Les eines de desenvolupament incorporen unes plantilles que faciliten la tasca a l'editor. La figura 2 mostra un exemple de dues malalties i les respectives *infoboxes*.

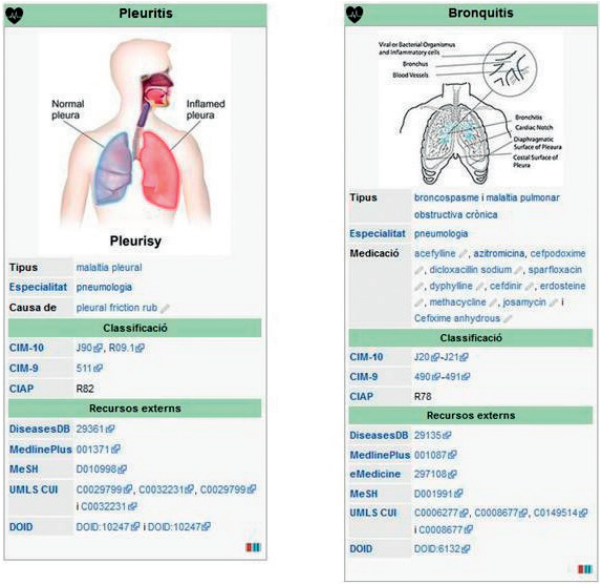

FIGURA 2. Exemples d'*infoboxes* en medicina.
FONT: Viquipèdia.

3. CREDIBILITAT DE LA VIQUIPÈDIA

Des de la seva creació el 2001, s'han dut a terme nombrosos treballs per estudiar i valorar els continguts de la Viquipèdia.[8] La majoria d'ells la comparen amb altres fonts tradicionals subjectes a revisió com ara l'*Encyclopædia Britannica* o *Encarta*. En general, els resultats indiquen que la qualitat dels articles és comparable a la que es troba en les enciclopèdies tradicionals però variable en relació amb recursos especialitzats.

Els editors disposen d'eines que els permeten veure la història de cada article i la corresponent pàgina de discussió. Els usuaris, en canvi, malgrat comptar amb accés a aquesta pàgina, no solen fer-ho per una qüestió de temps i també perquè se'ls pot fer difícil comprovar fins a quin punt la informació d'un article determinat és fiable. Com a conseqüència, l'usuari s'exposa a articles de qualitat variable.

S'ha de tenir en compte que la Fundació Wikimedia garanteix que tots els articles són revisats per experts de totes les branques de la ciència, incloent-hi els articles amb un alt contingut tècnic. Cada idioma té els seus propis mecanismes de control de qualitat. Els millors solen portar una marca de qualitat que es concreta en el fet que pertanyen a una categoria específica: «featured articles», en anglès; «artí-

8. Per a més informació, podeu consultar la pàgina *https://en.wikipedia.org/wiki/Reliability_of_Wikipedia*.

culos destacados», en espanyol, i «articles de qualitat», en català.[9] Hi ha una sèrie de criteris que ha de complir un article per ser considerat d'aquest tipus: que estigui ben redactat, que sigui complet, neutral, etc. Aquesta categoria sol ser oculta, és a dir, no està visible com les altres categories que tenen associades totes les pàgines.

Com ja s'ha comentat, l'origen de la informació de la Viquipèdia no és l'habitual en aquest tipus de recursos; és a dir, la producció per part d'un grup d'experts. Potencialment, la informació pot ésser afegida per qualsevol persona que compleixi unes normes d'edició establertes (i actualitzades a mesura que se'n veu la necessitat). Això pot crear i crea dubtes seriosos entre els experts sobre la fiabilitat de la informació que es troba en aquest recurs.

Per exemple, en [5] s'afirma que els articles sobre malalties cardiovasculars presenten errors per omissió.[10] D'altra banda, en [6] s'ha fet un estudi per tal d'identificar les tendències en l'ús de la WP com una referència en publicacions científiques amb comitè de revisió entre els anys 2002 i 2015. La conclusió és que troben citacions a la WP en revistes d'impacte i en articles produïts per acadèmics d'institucions rellevants.

4. Accessibilitat

La manera d'accedir a la Viquipèdia més coneguda i àmpliament utilitzada és mitjançant un navegador web.[11] Tota la informació que es mostra en una pàgina qualsevol (vegeu l'apartat 2) de la WP està formatada com una pàgina web tradicional.[12] Però la seva utilitat no seria completa si no fos possible accedir-hi també d'una forma automatitzada. Només d'aquesta manera el contingut d'aquest recurs pot ésser utilitzat, per exemple, per les aplicacions típiques del PLN.

L'extracció del text contingut en una pàgina de la WP pot fer-se molt fàcilment amb un buscador web i un *parser*. L'estructura regular de les pàgines permet la utilització d'aquest procediment. Tot i això, aquesta tècnica no sempre és satisfactòria, en particular si es vol augmentar el ventall d'aplicacions.

Afortunadament, hi ha múltiples aplicacions informàtiques que faciliten l'accés a la Viquipèdia. Existeixen llibreries que permeten accedir-hi utilitzant di-

9. En la WP en català hi ha 793 pàgines que es consideren articles de qualitat. Consulteu la pàgina «articles de qualitat» per saber-ne més detalls.

10. Cal destacar que aquestes mancances no fan referència a l'existència de la pàgina mateixa sinó al contingut de l'article. Es troben a faltar figures i taules per clarificar algunes pàgines; també hi ha deficiències pel que fa referència a la fisiopatologia, els mecanismes, l'enfocament diagnòstic i el pla de gestió.

11. Segons la companyia britànica YouGov, la WP és el setè lloc web més popular al Regne Unit. Segons les estadístiques proporcionades per Wikimedia Statistics, en l'últim any, les pàgines en català han tingut 210,78 milions de visualitzacions a tot el món.

12. En realitat, el format utilitzat no és l'HTML com en una pàgina web convencional, sinó un de molt semblant que s'anomena *wiki markup language*.

versos llenguatges de programació (Python, Perl, Java, Javascript, etc.) tant des d'estacions de treball com des de dispositius mòbils.

Malgrat les formes d'accés a la Viquipèdia que acabem d'esmentar, si es necessita un accés repetitiu i àgil, aquests mètodes no són eficients. En [3] s'aborda aquest problema des d'un altre punt de vista. Els autors proposen la creació d'un programari que permet convertir la descàrrega completa (o *dump*) d'aquest recurs al format SQL[13] i carregar-lo posteriorment a una base de dades convencional. D'aquesta manera, el temps d'accés a qualsevol consulta disminueix, fent possible l'ús intensiu d'aquest recurs com el que es necessita en les aplicacions que es descriuen en l'apartat 6. L'avantatge que representa la velocitat d'accés es veu mitigada per la necessitat de reproduir el procés de descàrrega i conversió a base de dades cada vegada que es considera necessari actualitzar el recurs.

Una altra opció és la utilització d'eines com ara el Wikipedia Miner Toolkit [7], un programari de lliure disposició que permet integrar l'accés a la WP en aplicacions pròpies. La idea és compartir algoritmes i codi en lloc de recursos. Inclou, entre altres coses, una API de Java que permet accedir i explorar les categories, pàgines i redireccions de la WP. S'hi inclouen també programaris per al processament dels *dumps*, mesures de similitud entre pàgines, serveis web, etc.

DBpedia [8] és un projecte col·laboratiu per a l'extracció d'informació estructurada i multilingüe de la WP, Viquidata i Viquimèdia Commons per fer-la accessible lliurement en la web utilitzant les tecnologies de la web semàntica i dades enllaçades.[14] La informació s'emmagatzema mitjançant l'estàndard RDF[15] mentre que per a les consultes a la base de dades s'utilitza l'SPARQL.[16] Des de fa pocs anys, aquestes tecnologies permeten, a través de llibreries específiques i en diversos llenguatges de programació, una altra forma d'accés molt ràpid i eficient a la Viquipèdia (juntament amb d'altres recursos que complementen la informació disponible).

5. Utilització de la Viquipèdia

Els diferents mètodes per accedir a la Viquipèdia que s'han mostrat en l'apartat 4 permeten un accés àgil i eficaç per al desenvolupament d'importants projec-

13. El contingut de la WP i altres recursos relacionats estan disponibles en format XML i poden descarregar-se lliurement des de *http://download.wikipedia.org*. Aquest recurs se sol actualitzar almenys un cop al mes.

14. Per al català només existeix una versió en preparació a *http://ca.dbpedia.org*.

15. L'RDF és un marc per a la representació de recursos a la web que s'ha dissenyat per ser utilitzat exclusivament entre ordinadors. Utilitza l'XML i forma part de la W3C's Semantic Web Activity.

16. L'SPARQL és un llenguatge d'interrogació per a grafs RDF i un protocol per accedir a aquests grafs dissenyat pel W3C RDF Data Access Working Group. Un graf RDF és un conjunt de tripletes. Una tripleta consisteix en un subjecte, un predicat i un objecte que conformen un fet complet. Un conjunt de tripletes enllaçades forma un graf de coneixement o graf RDF.

tes com els que es mencionaran en l'apartat 6. De totes maneres, és important tenir en compte les qüestions següents:

— Tipus de graf. El graf de categories no és una taxonomia, encara que és molt convenient i útil considerar-lo com a tal. La denominació mateixa de les relacions entre categories ja dona a entendre que aquest enllaç no sempre indica una relació d'hiponímia/hiponímia. Com ja s'ha mencionat en l'apartat 2, existeixen les categories de servei, que s'utilitzen per gestionar i estructurar el recurs o bé d'altres de caràcter enciclopèdic. Per exemple: agrupar esborranys; articles incomplets o que necessiten revisió; classificació de temes per any, país, regió, etc. Aquest fet no es pot deixar de tenir en compte quan s'explora aquest graf.

— Atribució de categoria a pàgines. Aquesta assignació pot respondre més a criteris enciclopèdics que taxonòmics. Per exemple, certes categories poden ésser assignades per establir informació de tipus (p. ex., «Enrico Fermi» → «físics teòrics»), de nacionalitat (p. ex., «Pau Casals» → «violoncel·listes catalans»), activitat («José de San Martín» → «militars argentins»), etc.

— Circularitat. Ambdós grafs presenten el problema de la formació de cicles. La figura 3 mostra un exemple real on es veu com les categories «ciències socials», «sociologia» i «societat» formen un cicle. La conseqüència és que, si no es prenen les precaucions oportunes, es bloqueja el recorregut d'una part del graf.

— Enllaços entre pàgines. Els enllaços entre alguns mots d'un article i una altra pàgina no tenen cap significat semàntic definit encara que el criteri d'edició sigui que han de ser rellevants per comprendre l'article. Alguns d'ells són superflus, a vegades erronis, o, fins i tot, poden referir-se a pàgines que encara no existeixen.

— Considerem, per exemple, la pàgina de «Galileo Galilei». Hi ha un enllaç amb la pàgina «països de parla catalana» que només serveix al lector per indicar-li en quin lloc geogràfic aquest científic es coneix com a «Galileu». Altres vegades, l'enllaç és manifestament inadequat. Considerem, per exemple, la pàgina d'«objecte»; la primera frase diu: «Un objecte és un ens limitat amb una funció precisa i…». La paraula «funció» té un enllaç cap a la pàgina de «funció» en el sentit utilitzat en matemàtiques però no en el sentit apropiat en aquest article. Aquestes circumstàncies dificulten la utilització d'aquest graf.

— Ubicació de les categories. La posició d'algunes categories pot causar un cert desconcert i crítiques entre els especialistes del domini. Considerem el cas de «medicina» i, dintre d'aquest domini, la categoria «veterinària». En anglès existeix la categoria *veterinary medicine* com a subcategoria de *medicine*, mentre que en català «medicina» i «veterinària» estan al mateix nivell i ambdues són subcategories de «ciències de la salut». Aquesta diferència pot ser polèmica i no és innòcua, ja que en aplicacions terminològiques pot provocar la selecció i/o validació de termes que no interessen per a l'àmbit de la «medicina». Aquest fet ens permet

recordar l'important paper que tenen els editors tant pel que fa a les pàgines com a les categories.

— Desenvolupament asimètric. El mecanisme d'ampliació d'aquest recurs fa que certes àrees del coneixement es desenvolupin més que altres.

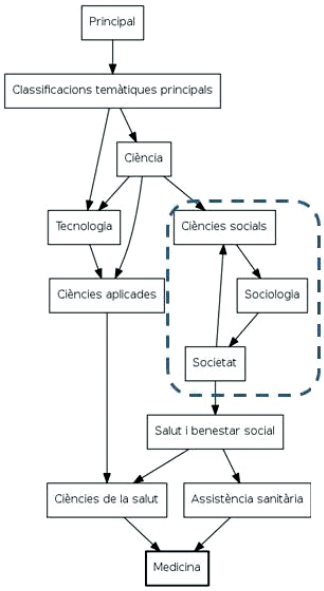

FIGURA 3. Exemple de cicles en l'estructura de categories de la Viquipèdia.
FONT: Elaboració pròpia.

Una qüestió important a tenir en compte és que aquest recurs pot ésser utilitzat també en altres llengües. Un requisit fonamental perquè això sigui possible, en la llengua que s'està considerant, és el grau de compleció de la WP i l'existència de recursos bàsics per al PLN raonablement complets.

6. APLICACIONS DE LA VIQUIPÈDIA

Des de la seva creació i malgrat les qüestions que hem plantejat en els apartats 4 i 5, la WP s'ha utilitzat i s'utilitza sovint per a consultes típicament enciclopèdiques però també ha sigut objecte de molts altres usos menys coneguts. Aquests van des de tasques que li són pròpies (com ara la detecció automàtica d'errors o l'ampliació automàtica del contingut, entre d'altres) fins a tasques relacionades amb el PLN. En els subapartats següents presentem algunes de les aplicacions més rellevants en aquest últim àmbit.

6.1. *Millora o estudi de la Viquipèdia mateixa*

Existeixen molts treballs que es dediquen a estudiar la qualitat dels articles. En [9], per exemple, es pren en consideració un centenar d'atributs lingüístics[17] per estudiar la qualitat de la WP en polonès. Aquest estudi s'ha fet sobre la base de 500.000 articles escollits aleatòriament, i el resultat és que se'n poden considerar de qualitat entre el 4 % i el 5 %.

També s'han desenvolupat programaris específics per a la detecció d'articles inadequats o vandàlics. En [10], es proposa un sistema motivat lingüísticament per a la detecció automàtica d'articles d'aquest tipus. Es fa palès que aquests articles tenen un estil propi i que es poden detectar mitjançant gramàtiques probabilístiques. En [11], es proposa un sistema amb objectius idèntics basat, però, en l'establiment de patrons i utilitzant un sistema d'aprenentatge automàtic.

Finalment, cal destacar els estudis fets dintre del domini mèdic. En primer lloc, en el treball de [12] s'estudia el temps que es triga a actualitzar la referència a articles d'impacte en aquest domini. La conclusió és que transcorre una mitjana de noranta dies a aparèixer citacions en publicacions rellevants com ara: *Cochrane Database of Systematic Reviews*, *Nature* o *The Lancet*, entre altres. En segon lloc, destaquem el treball de [13], en què es descriu el *WikiProject Medicine* i els seus punts forts i febles, i se'l compara amb projectes semblants. Es destaca la importància d'aquest recurs per al públic en general, així com per a estudiants i professionals de la salut. Finalment, destaquem el treball descrit a [14], en què s'utilitza la WP per enriquir un glossari de termes radiològics per a usuaris no especialistes. Del total de 4.090 conceptes presents al glossari, 3.063 (el 74,9 %) han trobat una correspondència a una pàgina de la WP. A més a més, de 800 conceptes escollits aleatòriament, el 51 % s'han enriquit semiautomàticament amb imatges preses de la WP.

6.2. *Tasques pròpies del processament de la llengua*

En una primera aproximació al gran ventall d'aplicacions de la WP, cal observar la freqüència amb què apareix en els articles presentats als congressos que organitza l'Association for Computational Linguistics (ACL).[18] Les actes d'aquests esdeveniments es recullen en una base de dades que actualment conté més de 48.000 articles. En aquest recull es poden fer cerques mitjançant una interfície web[19] en què es pot veure que els articles que d'una manera o altra mencionen la WP es compten per centenars.

17. Per exemple: nom i tipus de nom, adjectiu, verb i tipus de verb, cas, freqüència, gènere, etc.

18. L'ACL és una organització de reconegut prestigi en l'àmbit de la lingüística computacional que organitza regularment un gran nombre de congressos i tallers.

19. *https://www.aclweb.org/anthology*.

Fem, a continuació, un breu recorregut per algunes de les aplicacions més comunes que utilitzen aquest recurs dintre del PLN:

— Creació de corpus monolingües. Un corpus és una col·lecció organitzada de textos i és un requisit bàsic per a qualsevol tasca relacionada amb la lingüística de corpus, el PLN, etc. Un corpus pot ser monolingüe o plurilingüe i, en aquest últim cas, els textos que el formen poden ser paral·lels o comparables. La Viquipèdia és un conjunt de textos ben formats que constitueixen un recurs ideal per a la constitució de corpus textuals en diferents llengües i temàtiques. Com ja s'ha mencionat, els articles estan formatats en una variant del llenguatge HTML. Per tant, l'únic requeriment és l'eliminació de les marques d'aquest tipus. Existeixen força exemples de corpus compilats a partir de la Viquipèdia en moltes llengües, entre les quals el català.[20] També trobem aquest recurs formant part de corpus més grans, com ara Linguatools,[21] Sketch Engine[22] o OPUS,[23] entre altres.

— Traducció automàtica. Com ja s'ha comentat en l'apartat 2, alguns articles en una llengua poden tenir l'article corresponent en una altra llengua. En conseqüència, es poden considerar com a textos comparables i s'utilitzen com un recurs per aquests sistemes.

D'aquesta manera, es creen models per tal d'extreure, per exemple, les frases paral·leles i utilitzar-les per entrenar un sistema d'aquest tipus. Aquests treballs poden referir-se a un àmbit general però també a dominis específics.

En [15] s'extreuen les frases per a les parelles castellà-anglès, alemany-anglès i romanès-anglès i s'entrena un sistema de traducció automàtica estadístic. Un treball similar es descriu a [16]. El treball s'ha fet per a l'anglès i el francès i s'ha restringit a temes relacionats amb els Alps.

— Resum automàtic multidocument (extractiu). Aquestes eines identifiquen, amb l'ajuda de la WP, els conceptes rellevants d'un document i les frases que els contenen. Aquestes són classificades d'acord amb la importància dels conceptes que inclouen. El sistema selecciona les frases més rellevants i les comprimeix per formar el resum [17].

— Creació de taxonomies multilingües. L'objectiu aquí és la integració de diverses edicions de la WP per formar una única taxonomia [18].

— Creació d'una ontologia de domini. En [19] es proposa la creació d'una ontologia de domini en dos passos. En primer lloc es crea automàticament una ontologia utilitzant la Viquipèdia mitjançant l'ús de les relacions implícites en les

20. Disponible a *https://repositori.upf.edu/handle/10230/20050*. Aquest corpus està format per un total de 390.000 articles amb 125,6 M de paraules i ha sigut processat amb les eines de l'Institut de Lingüística Aplicada de la Universitat Pompeu Fabra (IULA).

21. *https://linguatools.org/tools/corpora/*.

22. *https://www.sketchengine.eu/*.

23. *http://opus.nlpl.eu/*.

plantilles de les pàgines, les categories i les *infoboxes*. Finalment, es proposa una manera d'aprofitar la informació inicial i un cercador per completar l'ontologia.

— Creació de bases de coneixement: la WP és la font principal d'informació per a la creació dels recursos següents: DBpedia,[24] Freebase,[25] WikiNet[26] i YAGO.[27]

— Enllaç d'entitats (*entity linking* o *Wikify*). Aquesta tasca consisteix a enriquir algunes paraules d'un text amb enllaços a pàgines de la Viquipèdia. Es pot considerar com un tipus d'etiquetatge semàntic en què l'etiquetari és el conjunt de pàgines de la WP. Per tant, s'han d'identificar els termes rellevants del text i anotar-los de manera no ambigua amb la pàgina pertinent. En relació amb aquesta tasca destaquem les propostes presentades en [20], [21] i [22].

— Anàlisi semàntica. ESA (*explicit semantic analysis*) és una representació vectorial de text (paraules o documents) que utilitza un corpus com una base de coneixement. Aquest treball utilitza cada article de la WP com una unitat de coneixement; es va iniciar amb [23] i [24], i ha sigut objecte de múltiples millores i adaptacions. El camp d'aplicació d'aquesta tècnica és el càlcul de la similitud entre paraules, frases o documents.

— Càlcul de la relació semàntica entre paraules o textos. Es tracta d'utilitzar el coneixement implícit als enllaços entre pàgines, així com entre pàgines i categories. Vegeu l'exemple descrit a [25].

— Reconeixement i classificació d'entitats en diversos idiomes. En [26] es proposa utilitzar les anotacions de la WP com les anotacions inicials necessàries en tot sistema d'aprenentatge automàtic. Per aconseguir aquest objectiu parteix de les anotacions en una llengua i, utilitzant l'estructura bigraf de la WP, es troben les anotacions en la llengua destí. L'anotació aconseguida és suficient per poder ampliar-la aplicant mecanismes d'aprenentatge automàtic.

6.3. *Terminologia*

La WP és un recurs que fa explícit un ampli ventall de conceptes en forma de títols de pàgines i nom de categories. Podem concebre dues maneres d'explotació de tota aquesta informació per al treball terminològic segons es faci una exploració de dalt a baix (*top-down*) o bé de baix a dalt (*bottom-up*) en el graf de categories. En aquest apartat posarem dos exemples reals d'explotació utilitzant ambdós mecanismes d'exploració.

24. *https://wiki.dbpedia.org.*
25. Aquest recurs és inactiu des de 2016, quan va ésser integrat a Viquidata.
26. *https://www.h-its.org/software/wikinet-2/.*
27. *https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.*

La idea subjacent en els projectes que es descriuran és que es poden establir fronteres de domini en el graf de categories. És a dir, trobar categories per sota de les quals podem considerar, amb un cert grau de confiança, que totes les categories i les pàgines associades són pertinents en un cert domini.

Una vegada establertes aquestes fronteres, segons el tipus d'exploració que es faci podrem:

1. obtenir tots els termes d'un domini que estan registrats en aquest recurs (exploració *top-down*) o bé
2. determinar si un mot trobat a la Viquipèdia (com a pàgina o categoria) és o no pertinent en el domini d'interès (exploració *bottom-up*).

En una primera aproximació, l'establiment d'aquestes fronteres és molt simple: el nom del domini (o la categoria de la WP més propera) és la frontera cercada. Aquesta solució, però, no sempre és suficient; pot ésser necessari un treball d'exploració per afegir altres categories que la complementin. Un exemple d'aquesta situació es dona quan volem establir les fronteres per a la medicina. En una aproximació podríem escollir la categoria «medicina». Però una anàlisi del graf de categories ens mostra que les seves supercategories són: «ciències de la salut» i «assistència sanitària». Una anàlisi d'aquestes dues categories ens podria portar a acceptar la primera i a descartar la segona. En el primer cas s'inclouen subcategories com ara «infermeria», «nutrició» i «odontologia», entre d'altres. Mentre que la categoria «assistència sanitària» inclou subcategories com ara «metges asiàtics», «sanitat per país» i «organitzacions sanitàries», entre d'altres, la gran majoria de les quals no són d'interès terminològic. No obstant això, seria convenient una revisió acurada d'aquestes categories per si es consideren rellevants per a la tasca que ens proposem dur a terme. En qualsevol cas, l'exploració ha de prendre les precaucions oportunes per evitar considerar aquestes categories (i probablement les pàgines que en depenen) com a terminològiques.

Un problema comú a tot treball terminològic, ja sigui de compilació o d'extracció, és l'avaluació del material obtingut. No existeix una solució per a aquest tema, ja que, d'una banda, l'avaluació manual per experts és inviable pel seu cost i, de l'altra, si hi intervé més d'un expert, sorgeix el problema de la diferència de criteris entre ells. Per a una avaluació automàtica fora necessari tenir una llista de referència. En el cas de l'extracció, aquestes llistes, si existeixen, difícilment són completes. En el cas de la compilació és el producte que volem obtenir. Per tant, cal recórrer a avaluacions parcials o indirectes.

En els dos subapartats següents analitzarem amb més detall les possibilitats d'ambdós mecanismes i els resultats que s'han obtingut.

6.3.1. Recull de la terminologia d'un àmbit

La compilació de la terminologia d'un àmbit és una tasca necessària per a moltes aplicacions de PLN. L'adaptació al domini dels recursos existents per processos com ara l'etiquetatge semàntic, l'extracció de relacions, l'etiquetatge de rols semàntic, el resum automàtic, la traducció automàtica i altres depenen en gran mesura del fet de disposar d'aquestes terminologies.

L'obtenció de la terminologia d'un domini és una tasca complexa. Es presenten dos problemes: en primer lloc, s'ha de disposar d'un corpus de textos ben construït i, en segon lloc, cal extreure els termes d'aquest corpus.

La compilació d'un corpus d'un domini qualsevol és una tasca costosa en temps i recursos. L'obtenció dels termes del corpus ja compilat també és problemàtica. El processament manual és una tasca inassolible, per la qual cosa s'ha de recórrer a mètodes automàtics. Existeixen diverses solucions per a aquest últim problema (en aquest mateix subapartat se'n presenta una) amb resultats variables. Una altra solució seria disposar d'un recull ja compilat amb els termes del domini d'interès. D'aquesta manera, el problema es reduiria a cercar aquests termes en el text que s'està processant.[28]

En [27] i [28] es mostren dues aproximacions per recollir automàticament els termes d'un domini. Ambdues propostes utilitzen com a punt de partida la Viquipèdia, ja sigui directament (en el primer cas) o indirectament, a través de DBpedia (en el segon). L'objectiu en la primera aproximació és obtenir tots els termes d'alguns dominis en castellà i anglès, mentre que, en la segona, l'objectiu és semblant però estenent la tasca a una col·lecció de dominis. També s'afegeix la possibilitat de trobar la correspondència entre els termes de les dues llengües. En la resta d'aquest apartat descriurem breument el procediment proposat a [27].

L'objectiu d'aquesta proposta és recopilar de les fonts lèxiques de cada domini i per a les dues llengües tants termes com sigui possible i cercar la correspondència entre ells. Per aconseguir aquest objectiu es disposa de: *a*) un conjunt d'etiquetes de domini, *b*) un parell de llengües i *c*) recursos de lèxics que cobreixen tots els dominis en les dues llengües.[29]

El sistema proposat utilitza aquests recursos lèxics:

1. Multilingual Central Repository (MCR).[30] Aquest recurs segueix el model proposat pel projecte EWN [29] per crear una base de dades lèxica multilingüe

28. Òbviament, es podria dir que difícilment un recull de termes ja compilat tindria tots els termes del domini. De totes maneres, també es pot dir que segons l'aplicació de què es tracti aquest podria ésser suficient.

29. D'aquí es pot desprendre una limitació del sistema: només recollirà els termes de domini que estiguin presents a la WP.

30. *http://adimen.si.ehu.es/web/mcr/*.

amb els *wordnets* per a diverses llengües europees. Els *wordnets* de cadascuna de les llengües s'estructuren com el WordNet (WN) [30] de Princeton.

2. Extended WordNet Domains (XWND). Aquest recurs neix a partir de Wordnet Domains (WND) [31] i fou ampliat en [32]. L'objectiu és assignar informació de domini als *synsets* de WN. Aquesta informació té la forma d'una taxonomia amb 164 dominis. XWND és un recurs lèxic en què els *synsets* de WN incorporen informació de domini com a WND, però aquest recurs assigna a cada *synset* una probabilitat de pertànyer a cada domini del conjunt de dominis.

3. Wikipedia (WP). Projecte objecte d'aquest treball. Vegeu-ne els detalls a l'apartat 2.

4. DBpedia. Projecte que extreu informació estructurada i multilingüe de la Viquipèdia per fer-la accessible lliurement en la web utilitzant les tecnologies de la web semàntica i dades enllaçades.

El procés global està esquematitzat en la figura 4, s'aplica iterativament a cada parella domini-llengua la seqüència de processament descrita a continuació:

— Pas 1. Per utilitzar XWND com a etiquetari semàntic és necessari un procés de normalització per tal que les assignacions de probabilitats de domini a cada *synset* de l'MCR siguin comparables.

— Pas 2. Per a cada domini i llengua s'utilitza l'MCR i XWND per identificar tots els *synsets* que tinguin una alta probabilitat de pertànyer al domini. Es busquen les variants d'aquests *synsets* en la WP. El conjunt d'aquestes categories és avaluat utilitzant la mateixa WP (vegeu el subapartat 6.3.2) per eliminar aquelles categories que es troben per sota d'un cert llindar.

— Pas 3. S'obtenen totes les categories principals de cada domini.

— Pas 4. Es filtren les categories principals utilitzant XWND, les supercategories i la distància al *top*.

— Pas 5. Les categories principals seleccionades s'expandeixen utilitzant els enllaços del tipus «subcategoria».

— Pas 6. S'aplica un conjunt de mesures específiques que filtren les categories obtingudes en el pas anterior.

— Pas 7. S'obtenen el conjunt de pàgines enllaçades amb el conjunt de categories.

— Pas 8. Es posa en marxa un procés iteratiu en què, a cada cicle, el conjunt de pàgines i categories es reforça o restringeix mútuament. El procés continua fins que no hi ha cap variació d'un cicle al següent. El resultat és el conjunt de pàgines i categories definitiu.

— Pas 9. S'apliquen una sèrie de filtres per millorar la qualitat dels resultats. El primer consisteix a aplicar un algoritme *page-rank* sobre una representació en forma de graf dirigit del conjunt de categories i pàgines per a cada domini-llengua. Els termes menys fiables són eliminats. A continuació es resol el problema

dels candidats que pertanyen a més d'un domini. També s'eliminen els casos de pàgina i categoria que es refereixen al mateix terme. Quan un terme apareix en singular i en plural s'elimina l'última forma.

— Pas 10. S'utilitza la DBpedia per obtenir, on sigui possible, els termes equivalents en l'altre idioma.

— Pas 11. Es procedeix a l'avaluació del resultat per a cada domini i llengua i a la creació dels fitxers en format OLIF.

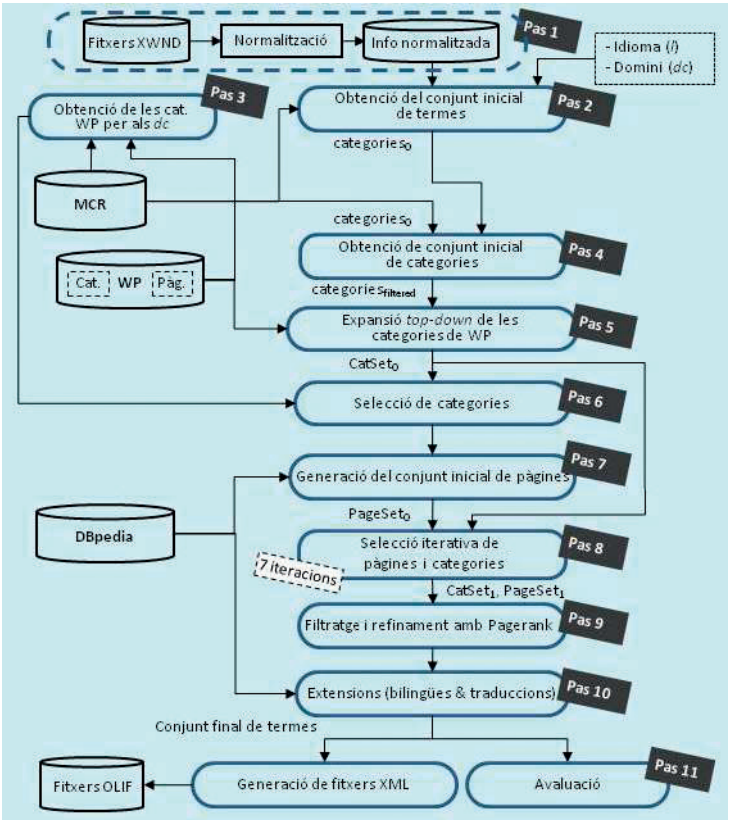

FIGURA 4. Esquema global per obtenir termes de tots els dominis inclosos en XWND.
FONT: Elaboració pròpia.

L'aplicació de la metodologia ja descrita ha permès obtenir 635.527 termes per als 164 dominis i les dues llengües. En la taula 1 es mostren els deu dominis en els quals s'han capturat més termes. L'última fila i l'última columna mostren les xifres globals. Per a cada domini es mostra el nombre de termes corresponent a

pàgines i categoria. S'inclou també el nombre de termes per als quals s'ha trobat una traducció.

Si es comparen aquests resultats amb el nombre de pàgines de la WP per a les llengües de treball,[31] els resultats quantitatius poden semblar escassos. S'ha de considerar, però, el caràcter enciclopèdic de la WP, fet que motiva la inclusió de moltes pàgines i categories que s'han de descartar perquè no es poden considerar terminològiques.

Taula 1. *Domini més freqüent per als termes obtinguts*

Domini	*Nombre de pàgines EN*	*Nombre de cat. EN*	*Nombre de pàgines ES*	*Nombre de cat. ES*	*Correspondència*	*Total*
social	41.710	2.230	6.206	808	6.583	50.954
free_time	26.819	461	1.223	136	716	28.639
animals	16.281	636	6.936	206	4.661	24.059
person	17.163	589	5.502	293	3.100	23.547
biology	13.847	754	4.318	339	3.036	19.258
medicine	13.353	852	4.227	423	3.473	18.855
plants	5.366	271	10.436	1.428	472	17.501
environment	14.124	901	2.105	235	1.996	17.365
sociology	13.715	1.315	1.874	452	1.738	17.356
industry	13.774	229	2.165	215	1.020	6.383
…						
Total	444.653	28.032	146.140	16.702	79.946	635.527

Font: Elaboració pròpia.

La col·lecció completa de tots els termes per a tots els dominis definits a WND es guarda en fitxers[32] amb el format definit per l'estàndard OLIF.[33] A tall d'exemple, la figura 5 mostra el contingut per al concepte «med_4434», que en

31. Segons les dades obtingudes a *https://en.wikipedia.org/wiki/List_of_Wikipedias* hi ha 5.845.347 pàgines en anglès i 1.516.327 pàgines en castellà.

32. La col·lecció completa serà disponible per descàrrega lliure.

33. *Open lexicon interchange format*. Es tracta d'un format lliure per a l'intercanvi d'informació lèxica i terminològica; vegeu-ne més informació a *http://www.olif.net/*.

anglès correspon a *assisted reproductive technology* i en castellà a *reproducción asistida*. Els dos termes es *mapegen* mútuament i, en conseqüència, comparteixen identificador. L'entrada inclou també altres informacions, com ara la categoria morfosintàctica, l'origen, el coeficient de fiabilitat, etc.

```
<entry ConceptUserId="med_4434">
 <mono>
  <keyDC>
   <canForm>assisted reproductive
     technology</canForm>
   <language>en</language>
   <ptOfSpeech>noun</ptOfSpeech>
  </keyDC>
 </mono>
 <monoDC>
  <monoAdmin>
   <confidence>0.801</confidence>
   <entrySource>WP page</entrySource>
  </monoAdmin>
 </monoDC>
</entry>
```

```
<entry ConceptUserId="med_4434">
 <mono>
  <keyDC>
   <canForm>reproducción asistida</canForm>
   <language>es</language>
   <ptOfSpeech>noun</ptOfSpeech>
  </keyDC>
 </mono>
 <monoDC>
  <monoAdmin>
   <confidence>0.800</confidence>
   <entrySource>WP page</entrySource>
  </monoAdmin>
 </monoDC>
</entry>
```

FIGURA 5. Exemple d'un terme en castellà i el seu equivalent en anglès en format OLIF.
FONT: Elaboració pròpia.

Pel que fa a l'avaluació dels termes obtinguts i per tal de minimitzar els problemes ja mencionats, s'han identificat escenaris diversos:

1. Avaluació parcial d'acord amb termes que apareixen tant a l'MCR com a la WP. Degut a la manca de fonts fiables de termes validats per la majoria dels dominis, el primer escenari consisteix a fer una avaluació restringida dels termes que apareixen tant a l'MCR com a la WP. En aquest cas, es donen per certes les assignacions fetes per XWND. Aquesta avaluació pot fer-se per a qualsevol parella domini-llengua.

En la figura 6 es mostren gràficament els termes trobats a l'MCR i la WP i es defineixen diferents grups de termes segons la seva pertinença. Aquesta avaluació es basa en el conjunt *C* perquè és el que agrupa els termes presents tant a l'MCR com a la WP i que pertanyen al domini. En la mateixa figura es mostren les fórmules de càlcul de precisió i cobertura.

2. Avaluació mitjançant fonts de domini externes. Es fa una avaluació completa per als dominis on existeixi una font fiable de referència. Aquest és el cas de medicina, que s'avalua mitjançant SNOMED-CT [33].

3. Avaluació mitjançant fonts externes. Per a l'idioma anglès i només per als termes inclosos en la Viquipèdia, es fa una comparació amb les assignacions fetes pel sistema de Niemann-Gurewych i descrit a [34] (NG).[34]

34. En [33] es proposa una alineació entre els sentits de WN anglès i les pàgines de la WP per a l'anglès.

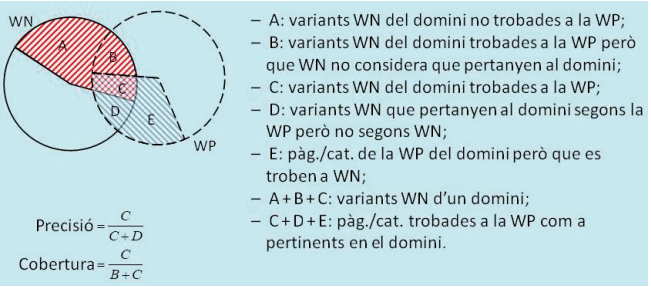

FIGURA 6. Definició dels conjunts de termes per a l'avaluació mitjançant l'MCR i la WP.
FONT: Elaboració pròpia.

4. Avaluació indirecta. L'últim escenari és indirecte i consisteix en l'ús del text dels articles de la WP associat als termes d'un domini per aprendre *word embeddings*[35] per a aquest domini i a continuació avaluar aquests *embeddings*.

L'avaluació es va limitar als dominis següents: agricultura, antropologia, arquitectura, medicina, música i turisme.

A continuació es mostren breument els resultats obtinguts amb els dos primers escenaris d'avaluació.

1. Avaluació parcial d'acord amb els termes que apareixen tant a WN com a la WP. Utilitzant el conjunt de termes segons s'ha definit en la figura 6, s'han calculat els valors de precisió i cobertura que es mostren en la taula 2. Per a cada idioma i domini es mostren el nombre inicial de termes a WN i els valors corresponents de precisió i cobertura.

2. Avaluació mitjançant fonts de domini externes. La utilització de SNOMED-CT permet, en principi, una avaluació millor dels termes de medicina, ja que es tracta d'una font reconeguda i molt utilitzada. De totes maneres, s'ha de tenir en compte que aquest recurs es defineix[36] com un «vocabulari de terminologia clínica utilitzat per professionals de la medicina per a l'intercanvi electrònic d'informació de salut». Per tant, no inclou tota la terminologia del domini sinó la que s'utilitza normalment en la pràctica diària. Aquest fet pot causar alguna indicació d'error falsa degut a:

a) Termes especialitzats. Algunes entrades es refereixen només a termes especialitzats. Vegeu, per exemple, el terme castellà *glándula*, que només existeix com a part d'un terme més específic com ara *glándula esofágica* o *glándula lagrimal*.

35. *Word embedding* és una tècnica molt utilitzada per a la representació del vocabulari d'un document. És capaç de capturar el context d'una paraula en un document, les similituds sintàctica i semàntica, la relació amb altres paraules, etc.

36. *https://searchhealthit.techtarget.com/definition/SNOMED-CT*.

TAULA 2. *Resultat de l'avaluació bàsica per a tots els dominis escollits*

Domini		*Turisme*		*Arquitectura*		*Música*	
Llengua		EN	ES	EN	ES	EN	ES
Termes a WN	Total (A+B+C)	556	180	219	30	1.121	234
	A la WP (C)	6	10	7	4	18	144
Precisió (%)		0,86	0,59	0,80	0,50	1,00	0,55
Cobertura (%)		1,00	1,00	0,89	1,00	1,00	1,00
Termes nous (D+E)		3.466	311	7.928	1.486	209	1.373
Total de termes nous al domini (C+D+E)		3.472	321	7.935	1.490	227	1.517

Domini		*Agricultura*		*Antropologia*		*Medicina*	
Llengua		EN	ES	EN	ES	EN	ES
Termes a WN	Total (A+B+C)	394	94	417	64	3.468	512
	A la WP (C)	4	6	11	15	499	196
Precisió (%)		0,67	0,55	0,79	0,54	0,78	0,55
Cobertura (%)		0,80	1,00	0,85	1,00	0,98	1,00
Termes nous (D+E)		394	510	2.294	1.040	13.282	2.523
Total de termes nous al domini (C+D+E)		398	516	2.395	1.055	13.881	2.719

NOTA: Els valors s'han d'interpretar d'acord amb la figura 6.
FONT: Elaboració pròpia.

b) Termes absents. Alguns termes relativament comuns estan presents a la WP però no en aquest recurs. Aquest fet és una de les causes de la baixa precisió mostrada en la taula 3.

c) Existència de termes complexos. S'inclouen com a termes simples alguns que en realitat són coordinats (p. ex., *enfermedades hereditarias y degenerativas del sistema nervioso central*).

En la taula 4 s'inclouen alguns exemples de termes validats i no validats, correctament i incorrectament.

TAULA 3. *Resultat de l'avaluació dels termes de l'àmbit de la medicina mitjançant SNOMED-CT*

	Llengua	
	EN	*ES*
Total de termes	13.382	2.523
Termes trobats	4.195	994
Precisió (%)	31,3	39,4

FONT: Elaboració pròpia.

TAULA 4. *Exemples de termes validats i no validats utilitzant SNOMED-CT*

	Llengua	
	Anglès	*Castellà*
Vàlid/*true*	*fibrosarcoma* *Kirschner wire* *pneumaturia*	*bronquiolitis* *duodeno* *placa dental*
Vàlid/*false*	*Alzheimer Research Forum* *Birmingham Accident Hospital*	*especialidades médicas*
No vàlid/*true*	*Rivalta test*	*otitis externa* *circulación portal hepática* *ortesis*
No vàlid/*false*	*high-throughput screening* *Kawasaki Medical School* *Goldwater rule*	*Condenado a vivir* *Asociación Pablo Ugarte* *huérfano del sida*

FONT: Elaboració pròpia.

Els resultats dels altres dos escenaris d'avaluació confirmen la validesa de la proposta que acabem de presentar i es poden consultar en [28].

6.3.2. Extracció de terminologia

La identificació dels termes presents en un text representa un coll d'ampolla per a la mineria de textos i, en conseqüència, és un tema de recerca important en l'àmbit del PLN. Podem veure l'extracció de termes com una tasca de marcatge semàntic per tal d'afegir al text informació sobre el significat. La manera d'abor-

dar aquesta tasca depèn dels recursos disponibles, principalment ontologies i llistes de termes. Si no es disposa d'aquesta informació cal recórrer a fonts d'informació indirecta de tipus lingüístic i/o estadístic. Els resultats que s'obtenen amb aquests mecanismes són limitats i per això aquestes eines tendeixen a afavorir la cobertura sobre la precisió. La conseqüència és que molts extractors obtenen llargues llistes de candidats que cal verificar manualment. Una de les raons d'aquest comportament és la manca d'informació semàntica. Les poques eines que utilitzen aquest tipus d'informació funcionen per a l'anglès. YATE [35] en constitueix una de les poques excepcions, ja que utilitza l'EWN com a font d'informació semàntica. Una altra possibilitat és utilitzar la WP com a font de coneixement. La WP representa una alternativa vàlida, i la utilització d'aquest recurs en aquest context és l'objectiu d'aquest apartat.

L'eina YATE està formada per diversos mòduls, un dels quals té com a funció determinar la terminologicitat d'un candidat utilitzant l'EWN. L'experiment que descriurem a continuació consisteix a construir un mòdul que tingui la mateixa funció utilitzant, però, la WP com a font de coneixement. L'àmbit escollit per a aquestes proves és la medicina, ja que disposa d'un cert nombre de fonts de coneixement que permeten superar les barreres comentades en el paràgraf anterior.[37] Per desenvolupar aquesta tasca s'haurà d'utilitzar l'estructura bigraf de la WP. A continuació veurem com utilitzar aquest recurs per calcular la terminologicitat d'un candidat.

Com ja s'ha comentat a l'inici d'aquest apartat, l'exploració dels grafs de la WP serà, en aquest cas, de baix cap a dalt; o sigui, a la inversa de l'exposada en el subapartat anterior. A partir del mot a valorar es procedeix a recórrer cap al *top* l'estructura bigraf fins a trobar una categoria que pugui considerar-se frontera de domini (FD) o bé arribar al *top*.

La figura 7 presenta de manera simplificada un fragment de l'estructura bigraf de la WP per tal de mostrar l'anàlisi del candidat a terme (CAT) *sang*. Aquesta figura mostra la situació que es presenta quan es vol analitzar el candidat *sang*. A la WP existeix la pàgina «sang», que té associada la categoria «sang». També es mostra la categoria «medicina», que és considerada com a frontera de domini. Podem imaginar una exploració del graf des de la pàgina amb el títol «sang» cap al *top*. És fàcil veure que existeixen múltiples camins possibles; alguns passen per la frontera de domini i altres no. Una possibilitat per valorar la terminologicitat o coeficient de domini (CD) del candidat *sang* seria considerar la relació entre aquest nombre de camins i que es concreta en l'equació, en què $NC_{frontera}(t)$ representa el nombre de camins al *top* que passen per la frontera i $NC_{total}(t)$, el total de camins al *top*.

37. Lamentablement, la majoria d'aquestes fonts són per a l'anglès, però n'hi ha d'altres (EWN) que es van poder ampliar en aquest àmbit en castellà i català.

$$CD_{nc}(t) = \frac{NC_{frontera}(t)}{NC_{total}(t)} \qquad \text{Equació 1}$$

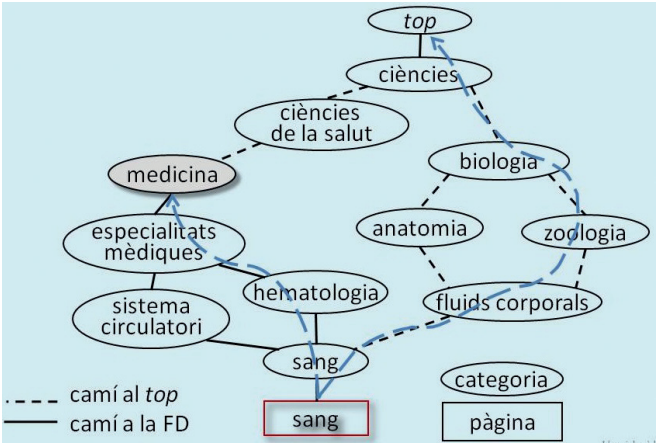

FIGURA 7. Exploració de l'estructura bigraf de la Viquipèdia per tal de valorar el candidat a terme *sang*.
FONT: Elaboració pròpia.

Utilitzant la mateixa figura 7, es poden identificar altres maneres de calcular la terminologicitat d'un candidat a terme. En les equacions 2 i 3 es mostren dues possibilitats més de càlcul. La primera es basa en la longitud dels camins (és a dir, el nombre de salts pàgina-categoria o categoria-categoria) i la segona, en la longitud mitjana dels camins.

$$CD_{lc}(t) = \frac{LC_{frontera}(t)}{LC_{total}(t)} \qquad \text{Equació 2}$$

en què $LC_{frontera}(t) =$ longitud dels camins a la frontera de domini
$LC_{total}(t) =$ longitud dels camins al *top*

$$CD_{lmc}(t) = \frac{LMC_{frontera}(t)}{LMC_{total}(t)} \qquad \text{Equació 3}$$

en què $LMC_{frontera}(t) =$ longitud mitjana dels camins a la frontera de domini
$LMC_{total}(t) =$ longitud mitjana dels camins al *top*

El resultat de l'aplicació d'aquests coeficients en l'avaluació d'un candidat a terme segons la informació de què disposa la WP pot ésser dividida en quatre grups:

1. CD..(t) = 1 → El candidat és un terme del domini;
2. 1 $>$ CD..(t) $>$ 0 → El candidat pot ésser utilitzat en diversos dominis. Normalment, com més gran és aquest valor més forta és la seva relació amb el domini;
3. CD..(t) = 0 → El candidat no pertany al domini;
4. CD..(t) = –1 → El candidat no està registrat a la WP i, en conseqüència, no se'n pot fer cap valoració.

A continuació es mostren els resultats que s'obtindrien si apliquéssim aquestes mesures al cas del candidat a terme *sang* utilitzant la figura 7:

$$CD_{nc}(t) = \frac{NC_{frontera}(t)}{NC_{total}(t)} = \frac{2}{2+2} = 0,5$$

$$CD_{lc}(t) = \frac{LC_{frontera}(t)}{LC_{total}(t)} = \frac{4+4}{6+6} = 0,66$$

$$CD_{lmc}(t) = \frac{LMC_{frontera}(t)}{LMC_{total}(t)} = \frac{4}{6} = 0,66$$

Aquest mecanisme té una limitació: només es poden valorar els termes que estan registrats com a pàgina o com a categoria. Una manera de superar aquesta barrera és utilitzar les pàgines de redirecció; en particular, les que fan referència a la redirecció dels adjectius relacionals.

Considerem, per exemple, el candidat a terme *maduració pulmonar*; a la WP no hi és i, per tant, no es pot reconèixer directament. Però si considerem que, aïlladament, ambdós mots es poden reconèixer com a termes en medicina podem també reconèixer el candidat sencer. En efecte, *maduració* té un coeficient de domini més gran que zero en medicina, mentre que *pulmonar* és un adjectiu relacional que la WP remet a *pulmó*, que també és reconegut com a terme. En conseqüència, assignem al terme complet un valor de terminologicitat que és una combinació del que obtenen cadascun dels mots separadament.

En [36] es presenta l'experiment que anticipàvem a l'inici d'aquest apartat i que permet comparar el comportament de YATE utilitzant l'EWN o bé la WP. A continuació, es presenta la metodologia emprada, els resultats obtinguts i la seva validació.

En la figura 8 es mostra l'esquema utilitzat per fer aquesta comparació. En aquest esquema es veu com el resultat del mòdul d'extracció de CAT de YATE s'utilitza per alimentar els dos mòduls d'anàlisi. A continuació, el resultat d'ambdós mòduls es compara amb el resultat de l'avaluació dels candidats del mateix text realitzada per especialistes del domini mèdic.

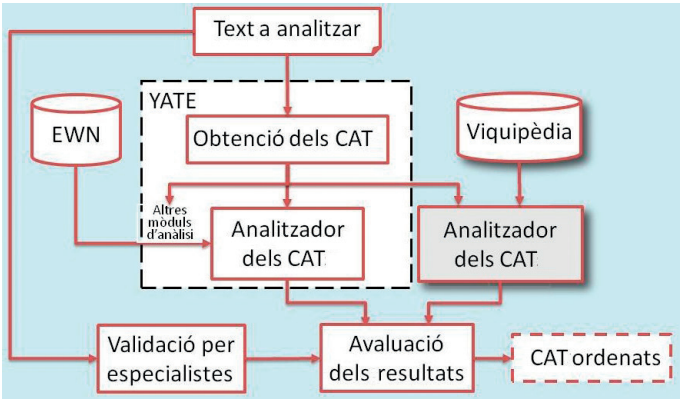

FIGURA 8. Esquema utilitzat per avaluar el mòdul de validació dels candidats a terme amb la Viquipèdia.
FONT: Elaboració pròpia.

Per realitzar l'estudi proposat s'han utilitzat textos del Corpus Tècnic de l'IULA [36] per a un total de 100 K paraules, aproximadament. Els textos van ser processats lingüísticament com és usual en PLN. L'avaluació s'ha efectuat amb les mesures de precisió i cobertura. Els candidats extrets per YATE s'han analitzat en primer lloc amb el mòdul de YATE (és a dir, utilitzant l'EWN com a font d'informació semàntica) i a continuació amb el mòdul que acabem de descriure (amb la WP i les tres mesures ja proposades). Cal mencionar que aquest extractor de termes només extreu els candidats que segueixen els patrons següents: *nom*, *nom-adjectiu* i *nom-preposició-nom*.[38] Els resultats obtinguts per a cada patró individualment, en termes de precisió i cobertura, es mostren en la figura 9, la figura 10 i la figura 11, mentre que en la figura 12 es mostren els resultats globals sense fer diferències entre patrons.

38. Es considera que aquests patrons cobreixen la gran majoria de termes, almenys en el domini de la medicina.

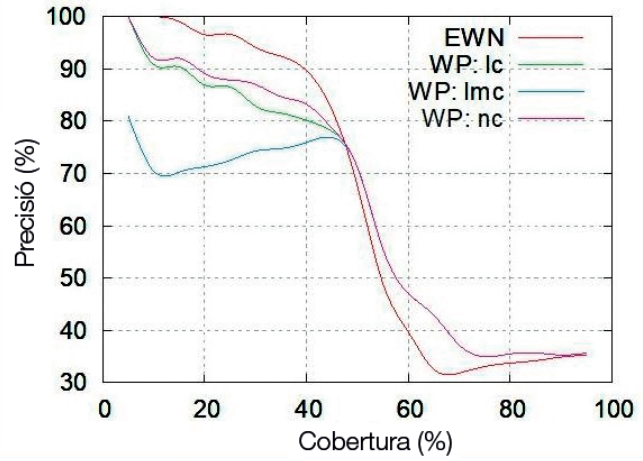

FIGURA 9. Avaluació de candidats a terme monolèxics nominals.
FONT: Elaboració pròpia.

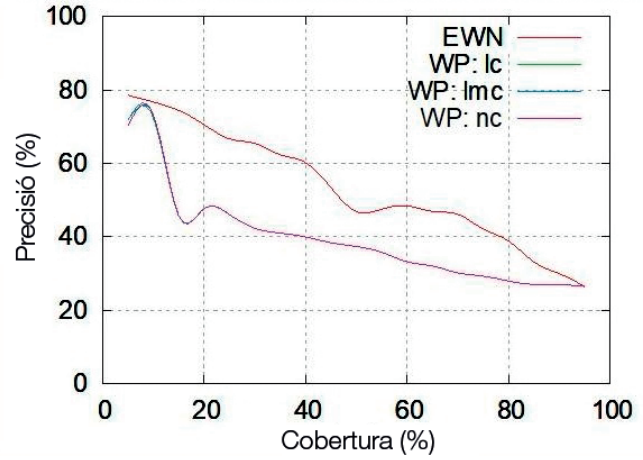

FIGURA 10. Avaluació de candidats a terme amb el patró *nom-adjectiu*.
FONT: Elaboració pròpia.

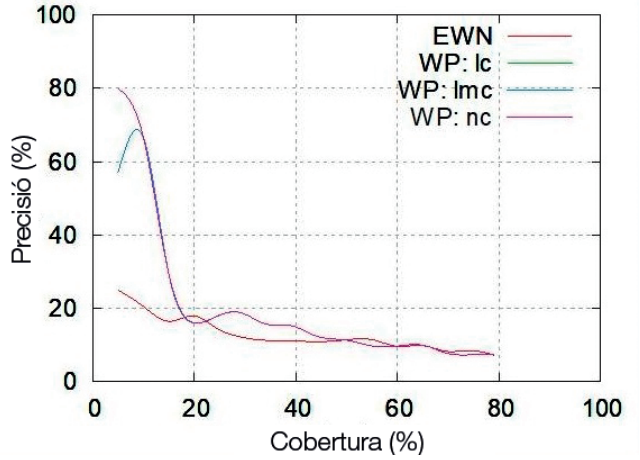

FIGURA 11. Avaluació de candidats a terme amb el patró *nom-preposició-nom*.
FONT: Elaboració pròpia.

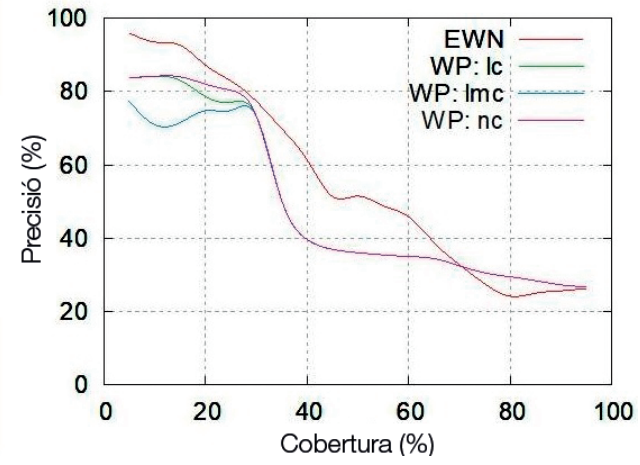

FIGURA 12. Avaluació de tots els candidats a terme conjuntament.
FONT: Elaboració pròpia.

Com es pot veure en la figura 9, la figura 10, la figura 11 i la figura 12, els resultats obtinguts utilitzant l'EWN són lleugerament superiors. Es considera que aquest comportament és, en part, degut al fet que la versió de l'EWN utilitzada està especialment adaptada per realitzar la tasca d'extracció de termes en medicina. De totes maneres, la diferència no és tan alta com podria esperar-se. Analitzem, a continuació, els resultats obtinguts per a cada patró individualment:

— Patró N: la diferència en els resultats obtinguts amb l'EWN i la WP varia entre un 10 % (CD_{nc}) i un 25 % (CD_{lmc}). Malgrat que aquesta diferència és important, cal mencionar que el coeficient CD_{nc} posiciona molt bé els CAT que estan inclosos a l'EWN. Cal mencionar també que hi ha CAT que existeixen a l'EWN però no a la WP.

— Patró NJ: en aquest cas el comportament de tots els CAT és molt similar i la diferència se situa al voltant del 25 %. Termes com ara *historia clínica* o *signo clínico* es classifiquen millor que amb l'EWN. Alguns termes són detectats amb la WP però no amb l'EWN (*poliarteritis nodosa*) i viceversa (*infección viral*).

— Patró NPN: en aquest cas tots els coeficients basats en la WP tenen un comportament millor que utilitzant l'EWN. La diferència és de prop del 10 % i la raó s'ha de buscar en el fet que l'EWN disposa de poques entrades per a aquest patró i l'estratègia de detecció no és gaire eficient. Al mateix temps, la WP incorpora moltes unitats amb aquest patró com, per exemple, *protocolo de tratamiento*, *grupo de riesgo* o *índice de mortalidad*, que reben la màxima qualificació amb la WP. En canvi, la qualificació d'aquests termes és molt reduïda amb l'EWN degut al fet que l'entrada completa no hi és i cada paraula per separat no té un significat especial en medicina. Cal mencionar també que, en el text analitzat, hi ha 910 candidats però només n'hi ha 39 a la WP i només 14 tenen un coeficient més gran que zero.

— Tots els patrons: en aquest comportament global la diferència es redueix un 5 % pel que fa a la precisió i un 30 % a la cobertura.

— Cal prestar una atenció especial al fet que, com ja s'ha comentat, la selecció de termes feta per especialistes és problemàtica. Com a exemple, podem dir que mots com ara *epitelio* o *medicina interna* s'han detectat correctament per ambdós sistemes però no s'han considerat com a termes pels especialistes; en conseqüència, el sistema d'avaluació els ha considerat errors.

En resum, podem considerar que l'extracció de termes en textos de biomedicina mitjançant la Viquipèdia pot ésser vàlida encara que la seva eficàcia pugui variar en funció del domini i del grau d'especialització del text a analitzar.

Aquesta estratègia per reconèixer termes ha sigut aplicada en diverses ocasions i àmbits:

— extracció de termes en medicina [36];
— estudi de la terminologia emprada en llibres de text a Mèxic: [38] i [39];
— projectes de màster en llengua italiana, portuguesa i francesa;
— projectes Alinea i APLE2 de l'IULA.[39]

39. A *http://eines.iula.upf.edu/WikiYATE/wikiYate.html* hi ha disponible una interfície gràfica que permet visualitzar els documents d'aquest projecte per a tots els àmbits del Corpus Tècnic de l'IULA i els termes escollits amb els contextos respectius.

7. Conclusions

En aquest treball s'han mostrat diferents aspectes de la Viquipèdia que van des d'una descripció detallada de l'estructura interna fins a la informació que conté. S'han analitzat algunes característiques tècniques que obliguen a prendre certes precaucions quan es consulta aquest recurs i, en particular, quan es vol utilitzar l'estructura bigraf. Aquestes qüestions no impedeixen que la Viquipèdia s'hagi utilitzat en multitud d'aplicacions per al PLN. Una qüestió sempre pendent quan es parla de la Viquipèdia és la credibilitat. El fet que no existeix un equip d'editors centralitzat fa que aquest tema surti freqüentment a la palestra. Al respecte, es mostren molts estudis i fins i tot propostes de millora i/o detecció de vandalisme en algunes pàgines. La Fundació Wikimedia és molt conscient d'aquest problema i la política de creació i edició de pàgines ha anat evolucionant a mesura que van apareixent manipulacions i altres problemes. S'han descrit algunes inconsistències quan analitzem la posició d'una mateixa categoria (cat.) en diferents llengües. Aquestes inconsistències no s'han estudiat com mereixen i poden representar un problema en funció de l'aplicació que s'estigui donant a la WP. A més a més, posa de relleu la importància del treball dels editors en la definició de quines categories s'assignen a una pàgina i de quin paper tenen aquestes categories en l'arbre de categories.

Malgrat el que acabem de mencionar, la Viquipèdia ha sigut molt utilitzada en diverses àrees del PLN. Al mateix temps, és part essencial d'altres fonts de coneixement de gran difusió que la prenen com a referència. En aquest treball es mencionen algunes d'aquestes aplicacions i, en particular, s'han descrit amb cert detall dues aplicacions específiques de l'àmbit de la terminologia, com són construir automàticament un recull de termes d'un domini i, donat un text, fer l'extracció de termes d'un domini. Aquests treballs demostren que és factible utilitzar la Viquipèdia en treballs terminològics. En ambdós casos queden alguns dubtes sobre el seu comportament quan l'àmbit de treball és molt especialitzat o quan som davant d'un àmbit transversal. En qualsevol cas, és un recurs que mereix ser tingut en compte en aplicacions que requereixen l'us de la terminologia.

Finalment, cal lamentar que la majoria dels treballs esmentats en l'apartat 6 facin referència fonamentalment a tasques fetes amb la llengua anglesa. El català, tot i ser una llengua molt present a l'univers Wikipedia, és clarament minoritari pel que fa als projectes que l'utilitzen.

Bibliografia

[1] Jullien, N. (2012). «What we know about Wikipedia: A review of the literature analyzing the project(s)». *HAL* [en línia]. 86 p. <https://hal.archives-ouvertes.fr/hal-00857208/document>.

[2] Nielsen, F. A. (2012). «Wikipedia research and tools: review and comments». *SSRN Electronic Journal* [en línia]. 66 p. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2129874>.
[3] Zesch, T.; Müller, C.; Gurevych, I. (2008). «Extracting lexical semantic knowledge from Wikipedia and Wiktionary». A: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. París: ELRA, p. 1646-1652.
[4] Kaptein, R.; Kamps, J. (2013). «Exploiting the category structure of Wikipedia for entity ranking». *Artificial Intelligence,* vol. 194, p. 111-129.
[5] Azer, S. [*et al.*] (2015). «Accuracy and readability of cardiovascular entries on Wikipedia: are they reliable learning resources for medical students?». *BMJ Open*, vol. 5 (10), p. 1-14. DOI: 10.1136/bmjopen-2015-008187.
[6] Tomaszewski, R.; MacDonald, K. I. (2016). «A study of citations to Wikipedia in scholarly publications». *Science & Technology Libraries*, vol. 35 (3), p. 246-261.
[7] Milne, D.; Witten, I. H. (2013). «An open-source toolkit for mining Wikipedia». *Artificial Intelligence*, vol. 194, p. 222-239.
[8] Lehmann, J. [*et al.*] (2015). «DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia». *Semantic Web*, vol. 6 (2), p. 167-195.
[9] Lewoniewski, W.; Wecel, K.; Abramowicz, W. (2018). «Determining quality of articles in polish Wikipedia based on linguistic features». A: *International Conference on Information and Software Technologies*.
[10] Harpalani, M. [*et al.*] (2011). «Language of vandalism: improving Wikipedia vandalism detection via stylometric analysis». A: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland: Association for Computational Linguistics, p. 83-88.
[11] Sarabadani, A.; Halfaker, A.; Taraborelli, D. (2017). «Building automated vandalism detection tools for Wikidata». A: *Proceedings of the 26th International Conference on World Wide Web Companion*. Cantó de Ginebra (Suïssa): International World Wide Web Conferences Steering Committee, p. 1647-1654.
[12] Jemielniak, D.; Masukume, G.; Wilamowski, M. (2019). «The most influential medical journals according to Wikipedia: quantitative analysis». *Journal of Medical Internet Research*, vol. 21 (1). També disponible en línia a: <https://www.jmir.org/2019/1/e11429/>.
[13] Heilman, J. M. [*et al.*] (2011). «Wikipedia: a key tool for global public health promotion». *Journal of Medical Internet Research*, vol. 13 (1). També disponible en línia a: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3221335/>.
[14] Martin-Carreras, T.; Kahn, C. E. (2019) «Integrating Wikipedia articles and images into an information resource for radiology patients». *Journal of Digital Imaging* (1 juny), vol. 32 (3), p. 349-353.
[15] Tufiş, D. [*et al.*] (2013). «Wikipedia as an SMT training corpus». A: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar: INCOMA, p. 702-709.

[16] Plamada, M.; Volk, M. (2013). «Mining for domain-specific parallel text from Wikipedia». A: *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*. Sofia: ACL, p. 112-120.

[17] Nastase, V.; Filippova, K.; Milne, D. (2009). «Summarizing with encyclopedic knowledge». A: *Proceedings of the 2nd Text Analysis Conference*. Gaithersburg: National Institute of Standards and Technology.

[18] Melo, G.; Weikum, G. (2010). «MENTA: inducing multilingual taxonomies from Wikipedia». A: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. Nova York: ACM, p. 1099-1108.

[19] Sedigheh, K.; Abolghasem, M. S. (2015). «Automatic construction of domain ontology using Wikipedia and enhancing it by Google search engine». *Journal of Information Systems and Telecommunication* [ACM], vol. 3 (4), p. 248-258.

[20] Mihalcea, R.; Csomai, A. (2007). «Wikify!: linking documents to encyclopedic knowledge». A: *CIKM'07: Proceedings of the sixteenth ACM Conference on Information and Knowledge Management*. Nova York: ACM, p. 233-242. També disponible en línia a: <https://dl.acm.org/doi/proceedings/10.1145/1321440>.

[21] Ferragina, P.; Scaiella, U. (2010). «TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities)». A: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. Nova York: ACM, p. 1625-1628.

[22] Wang, Z. [*et al.*] (2017). «Entity linking in queries using word, mention and entity joint embedding». A: Wang, Z.; Turhan, A. Y.; Wang, K.; Zhang, X. (ed.). *Semantic Technology*, vol. 10675. Cham: Springer. (Lecture Notes in Computer Science)

[23] Gabrilovich, E.; Markovitch, S. (2009). «Wikipedia-based semantic interpretation for natural language processing». *Journal of Artificial Intelligence Research*, vol. 34, p. 443-498.

[24] Gabrilovich, E.; Markovitch, S. (2007). «Computing semantic relatedness using Wikipedia-based explicit semantic analysis». A: *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (Índia). San Francisco: Morgan Kaufmann, p. 1606-1611.

[25] Yazdani, M.; Popescu-Belis, A. (2013). «Computing text semantic relatedness using the contents and links of a hypertext encyclopedia». *Artificial Intelligence*, vol. 194, p. 176-202.

[26] Fernandes, E. R. [*et al.*] (2016). «Using Wikipedia for cross-language named entity recognition». A: *Big Data Analytics in the Social and Ubiquitous Context*. Cham: Springer.

[27] Vivaldi, J.; Rodríguez, H. (2012). «Using Wikipedia for domain terms extraction». A: *Proceedings of CHAT 2012: The 2nd Workshop on the Creation; Harmonization and Application of Terminology Resources*. Linköping: Linköping University Electronic Press, p. 3-10.

[28] Vivaldi, J.; Rodríguez, H. (en premsa). «Automatically producing semantically tagged bilingual terminologies».

[29] Vossen, P. (1998). «The EuroWordNet Annual Report 1998». Amsterdam: Vrije Universiteit. [Document de treball]
[30] Fellbaum, C. (1999). «Wordnet: an electronic lexical database». *The Library Quarterly* [Chicago: The University of Chicago Press], vol. 69, p. 406-408.
[31] Magnini, B.; Cavaglià, G. (2000). «Integrating subject field codes into WordNet». A: *Proceedings of the Language Resources and Evaluation Conference (LREC 2000)*. Atenes: ELRA, p. 1413-1418.
[32] Bentivogli, L. [*et al.*] (2004) «Revising the Wordnet domains hierarchy: semantics, coverage and balancing». A: *Proceedings of the Workshop on Multilingual Linguistic Ressources*. Stroudsburg: Coling, p. 94-101.
[33] Spackman, K. A.; Campbell, K. E.; Côté, R. A. (1997). «SNOMED RT: a reference terminology for health care». A: *Proceedings of the AMIA Annual Fall Symposium*. Nashville: Hanley & Belfus, vol. 4, p. 640-644.
[34] Niemann, E.; Gurevych, I. (2011). «The people's web meets linguistic knowledge: automatic sense alignment of Wikipedia and WordNet». A: *Proceedings of the 9th International Conference on Computational Semantics*. Oxford: ACL, p. 205-214.
[35] Vivaldi, J. (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Tesi doctoral. Barcelona: Universitat Politècnica de Catalunya.
[36] Vivaldi, J.; Rodríguez, H. (2011). «Using Wikipedia for term extraction in the biomedical domain: first experience». *Procesamiento del Lenguaje Natural*, vol. 45, p. 251-254.
[37] Vivaldi, J. (2009). «Corpus and exploitation tool: IULACT and bwanaNet». A: *A Survey on Corpus-based Research. Proceedings of the I International Conference on Corpus Linguistics (CICL 2009)*. Múrcia: Universidad de Murcia, p. 224-239.
[38] Cabrera-Diego, L. A. [*et al.*] (2011). «Using Wikipedia to validate term candidates for the Mexican basic scientific vocabulary». A: *Proceedings of the First International Conference on Terminology, Languages, and Content Resources (LaRC)*. Seül, p. 76-85.
[39] Cabrera-Diego, L. A. [*et al.*] (2012). «Using Wikipedia to validate the terminology found in a corpus of basic textbooks». A: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: ELRA, p. 3820-3827.
[40] Vivaldi, J.; Rodríguez, H.; Rigau, G. (2013). «Combining Wikipedia and WordNet for improving domain terms compilation». A: *Proceedings of the 14th International Conference, CICLing 2013* (Samos). Berlín: Springer.
[41] Vivaldi, J.; Rodríguez, H. (2014). «Arabic medical terms compilation from Wikipedia». A: *2014 Third IEEE International Colloquium in Information Science and Technology (CIST)*. Tetuan: IEEE, p. 248-253.