WALLACE
CURS 2018-2019

En primer lloc, vull agrair a la meva tutora del Treball de Recerca la seva orientació i supervisió d'aquest projecte i la consolidació en l'última etapa. També agraeixo a la meva tutora inicial els seus consells inicials, els quals em van ajudar a encaminar el treball.

En segon lloc, vull especificar que la part pràctica d'aquest treball no hauria estat possible sense el guiatge, seguiment i mestratge de la doctora en Lingüística Computacional Montserrat Marimon Felipe, essencial en la creació i execució del programa analitzador de notícies.

Finalment, vull agrair el suport de Lluís Padró, creador i fundador de FreeLing*, en la secció d'*output *i extracció de resultats.*

ÍNDEX DE CONTINGUTS

1. Introducció

D'una banda, sempre m'he sentit atret per la Lingüística en tots els seus camps: la gramàtica, la semàntica, el lèxic, la pragmàtica, etc. De l'altra, considero la informàtica un aspecte molt important en els meus estudis actuals i futurs. El Processament del Llenguatge Natural (PLN), el tema del treball en qüestió, fusiona aquestes dues disciplines per donar lloc a un nou àmbit: la Lingüística Computacional (LC). Aquest nou sector està directament relacionat amb la Intel·ligència Artificial (IA) i té un potencial de recerca i desenvolupament immens.

És la comunió de les dues disciplines i els meus interessos acadèmics el que ha fet plantejar-me la possibilitat d'entrar al món del Processament del Llenguatge Natural (PLN) i de la Lingüística Computacional (LC).

L'objectiu de la lingüística és explicar com funciona el llenguatge; i el de la Lingüística Computacional és explicar-ho d'una forma representable en un ordinador. Així mateix, la LC és un camp molt extens i inclou molts subapartats de diferents dominis, la majoria dels quals es troben en constant desenvolupament -concretament les àrees relacionades amb la Intel·ligència Artificial -.

La recerca en aquest camp, tant teòrica com pràctica, em permetrà fer un tast del que podrien ser els meus estudis universitaris i, alhora, aprofundir en els coneixements que ja tinc de lingüística i iniciar-me en el llenguatge de la Intel·ligència Artificial. Per aquest motiu he titulat el treball: "Una aproximació al Processament del Llenguatge Natural".

1.1. Objectius

1.1.1. Específics

- Programar i exemplificar informàticament una de les aplicacions descrites a la part teòrica: la recuperació i l'extracció d'informació. Es tracta d'un *software* que analitza i processa articles periodístics i n'extreu conclusions en forma de dades estadístiques.
- Analitzar el punt de vista de dos investigadors que treballen i fan recerca en el món del Processament del Llenguatge Natural a partir de bases de formació totalment diferents: la lingüística i l'enginyeriainformàtica.

1.1.2. Generals

- Donar a conèixer la LC, un àmbit poc conegut que fusiona dues grans disciplines: la lingüística i la informàtica. Conscienciar el lector de la importància que obtindrà el Processament del Llenguatge Natural en un futur pròxim i l'ús que en fem a la nostra vidaquotidiana.
- Comprovar que el PNL és un camp d'investigació que m'atrau suficientment per enfocar els meus estudis i el meu futur cap aquest àmbit.

1.2. Metodologia

Per introduir-me al món de la Lingüística Computacional, he hagut de familiaritzar-me amb la terminologia pròpia d'aquesta ciència. Així doncs, he creat definicions i he après conceptes bàsics del tema en qüestió a partir de la consulta de pàgines *web*, articles digitals i treballs universitaris. Al llarg de l'elaboració del treball, el nivell del lèxic tècnic i específic s'ha anat ampliant, per la qual cosa la recerca de definicions, termes i conceptes ha sigut constant.

He de dir que el PNL és un camp relativament recent i la majoria de les fonts d'informació que he trobat són en anglès i, en alguns casos, en castellà. Aquest fet ha comportat una cerca lexicogràfica extensa per trobar vocabulari específic i tècnic en llengua catalana. No obstant això, moltes paraules i expressions angleses o bé no es poden traduir, o bé la seva traducció és poc habitual. Això ha provocat que en alguns punts del treball, com per exemple en l'apartat "Aplicacions", hi hagi inclòs tant la forma genuïna del concepte, com la traduïda.

A continuació, he redactat la part teòrica a partir d'estudis en què ja es parlava de lingüística i de programació, concretament de PNL: definició, objectius, origen i evolució, nivells d'anàlisi, mètodes i enfocaments i, finalment, aplicacions.

Per la part pràctica, volia contactar amb experts en Lingüística Computacional. I parlo d'experts, en plural, perquè tenia clar que volia conèixer dues persones que des de formacions ben diferents, una de "lletres" (Filologia Anglogermànica) i una altra de "ciències" (Enginyeria Informàtica), haguessin arribat al mateix punt professional: investigadors en el camp del Processament del Llenguatge Natural. És per això que he contactat amb la lingüista Montserrat Marimon i l'enginyer informàtic Lluís Padró, el currículum dels quals és remarcable. Ens hem pogut trobar diverses vegades, parlar

llargament i entrevistar-los formalment. De l'entrevista, n'he extreure informació imprescindible per la fase pràctica del treball.

Per la seva part, Lluís Padró m'ha presentat la plataforma *FreeLing,* creada per ell mateix, per dur a terme les anàlisis lingüístiques amb diferents nivells de complexitat.

Per altra part, he tingut l'oportunitat de treballar durant l'estiu al costat de la Doctora M. Marimon, la qual m'ha introduït en un llenguatge de programació nou per mi, el *Python*, necessari per la codificació i execució del programa informàtic que he utilitzat en l'elaboració d'aquest treball.

Paral·lelament, he fet una recerca de material textual (120 notícies, de 14 diaris diferents en català, castellà i anglès) per poder-hi aplicar les tècniques informàtiques apreses d'extracció, d'anàlisi i d'interpretació de dades. Aquesta part ha requerit una dedicació important de treball de recerca i científic d'assaig / error.

Finalment, un cop comprovat l'assoliment dels objectius, n'he redactat les conclusions.

2. Marc teòric: el Processament del Llenguatge Natural

2.1 Definició i objectiu

La definició de Processament del Llenguatge Natural (PLN) inclou diversos aspectes fonamentals de la lingüística i de la informàtica i, com que és un camp que encara es troba en desenvolupament, és complex i pot tenir diferents interpretacions. Una definició general podria ser:

*"*El Processament del Llenguatge Natural és un marc de motivació teòrica de tècniques computacionals que té com a objectiu analitzar i representar textos que incloguin un o més nivells d'anàlisi lingüística per aconseguir comprendre i processar el llenguatge com els humans per realitzar una sèrie de tasques o aplicacions.*"* [1]

Hi ha conceptes de la definició que són clau per entendre el PLN, com: tècniques computacionals, representació de textos en múltiples nivells d'anàlisi lingüística i sèrie de tasques o aplicacions. Aquests conceptes els presento i desenvolupo en les diferents seccions del treball.

En suma, el PLN és una branca de la Intel·ligència Artificial (IA) que intenta que els ordinadors processin el llenguatge humà. És a dir, l'objectiu del PLN és fer màquines que processin la llengua. Si la LC genera un model computable de la llengua, el PLN el podrà usar. Com veurem més endavant en l'apartat "Mètodes" es fan servir diversos enfocaments (estadístic, xarxes neuronals, etc.) per produir resultats útils, encara que el llenguatge no funciona realment així.

2.2 Història

2.2.1 Primera etapa (1940-1950)

El camp de la traducció automàtica va aparèixer per primer cop als anys 40 de la mà del matemàtic Warren Weaver i les seves idees relacionades amb la criptografia i la teoria de la informació per a la traducció d'idiomes com el rus i l'anglès a la Segona

[1] Liddy, E.D. 2001, *Natural Language Processing* in Encyclopedia of Library and Information Science.

Guerra Mundial. El 1949 va publicar un memoràndum de traducció, dissenyat per obtenir mètodes més eficients de traducció entre llengües que la forma simplista de "paraula per paraula", en la qual la freqüència d'errors era bastant elevada. Entre les seves propostes es destacaven els problemes de l'ambigüitat i els significats múltiples, i l'ús de mètodes criptogràfics aplicables a la traducció.

La intervenció de Noam Avram Chomsky -un dels lingüistes amb més influència en l'evolució de la gramàtica i el Processament del Llenguatge Natural- i la seva teoria que definia el llenguatge formal com una seqüència de símbols van ser claus pel desenvolupament del PNL i per l'aparició de noves hipòtesis.

2.2.2 Segona etapa (1957-1970)

En aquesta època es va realitzar una profunda recerca en l'àmbit dels enfocaments simbòlics (explicats en l'apartat "Mètodes"), repartida en dues línies d'investigació. La primera es va iniciar el 1957 amb la publicació *Syntactic Structures,* de Chomsky, i les teories de la gramàtica generativa. Aquestes defineixen la gramàtica com un sistema de regles i transformacions que generen totes les combinacions de mots que són considerades com a frases gramaticals d'una llengua, també impulsades per Chomsky.

La segona línia d'investigació era l'innovador camp de la Intel·ligència Artificial. Es van crear sistemes simples que tenien com a objectiu detectar la coincidència de patrons i la cerca de paraules clau en un text. D'aquesta manera van aparèixer els primers prototips de *question answering*, concepte explicat a l'apartat "Aplicacions".

2.2.3 Tercera etapa (1970-1993)

En aquestes dues dècades es va produir un gran desenvolupament per part dels nivells d'anàlisi semàntica i de discurs. Es van dissenyar sèries de programes de PLN enfocats en l'anàlisi i el processament del coneixement humà, com per exemple guions, plans i objectius. La base lògica i els paradigmes del PLN es van unir en sistemes que empraven la lògica de primer ordre per representar el nivell semàntic.

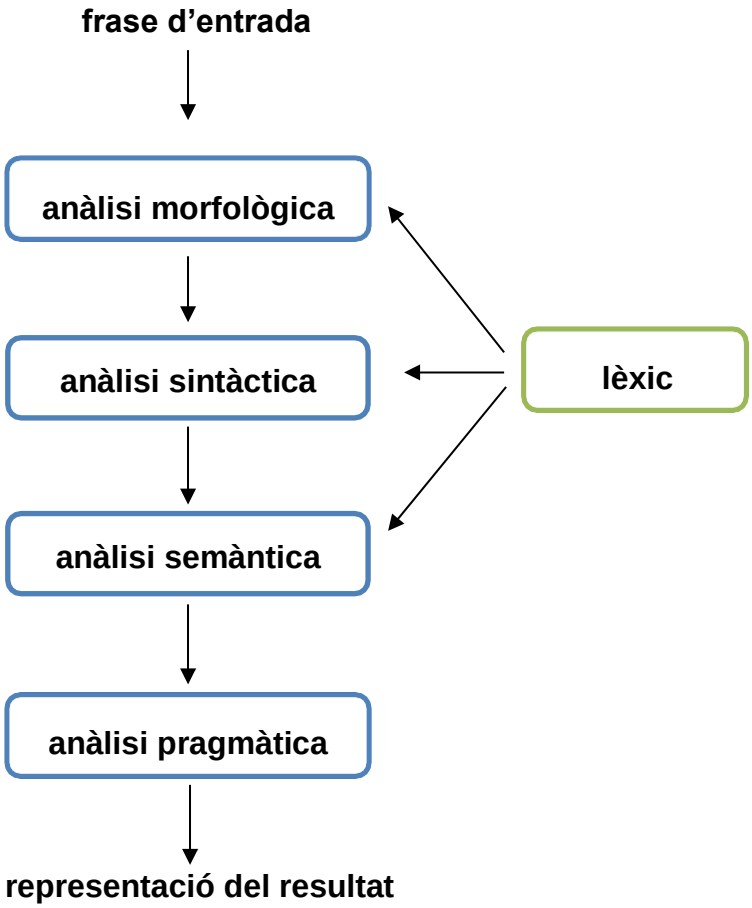
2.2.4 Quarta etapa (1993 fins a l'actualitat)

Aquesta època està marcada per l'estandardització dels models estadístics i de dades en el PLN. Els algoritmes d'anàlisi, l'etiquetatge de la categoria morfològica i el discurs van començar a incorporar les probabilitats i l'aprenentatge a la base de dades. També

van emprar estratègies d'avaluació a partir del reconeixement de la veu i l'extracció d'informació.

2.3 Nivells d'anàlisi

El mètode més efectiu per mostrar el procés que segueix el sistema de PLN és l'enfocament dels nivells de la llengua. Aquest mètode està compost per set nivells d'anàlisi diferents, tot i que n'hi ha quatre de fonamentals: el morfològic, el sintàctic, el semàntic i el pragmàtic. És un mètode seqüencial (com s'observa en la ***Il·lustració 1***) i el punt clau és la informació pròpia que aporta cada nivell, la seva funció i la importància dins d'una oració.



Il·lustració 1: *Els diferents nivells d'anàlisi del PLN ordenats de manera seqüencial per formar un procés d'anàlisi. FONT: elaboració pròpia.*

2.3.1 Nivell morfològic

Aquest nivell d'anàlisi fa referència a la naturalesa de la formació de les paraules, les quals estan compostes per morfemes -les unitats mínimes de significat-. El significat dels morfemes és el mateix encara que la paraula variï, per tant podem desglossar una paraula desconeguda en prefixos, sufixos i infixos i trobar-ne el significat. D'aquesta manera, el PNL també pot reconèixer el significat de cada morfema per obtenir el significat de les paraules.

La paraula "infelicitat", per exemple, està composta per tres morfemes: "in" significa "no" o "oposició" i "itat" significa "qualitat de". "Feliç" és un morfema independent (lexema), ja que pot aparèixer sol en forma de paraula. Els altres dos, en canvi, són morfemes dependents i s'ajunten als independents per modificar-los-hi el significat i la categoria gramatical, però no són paraules.

En la ***Il·lustració 2*** es mostra el resultat d'una anàlisi morfològica completa, en la qual s'assigna a cada paraula totes les seves possibles categories gramaticals i els trets morfològics.

La	**justícia**	**alemanya**	**descarta**	**el**	**delicte**	**de**	**rebel·lió**	**.**
el	justícia	alemany	descartar	el	delicte	de	rebel·lió	.
DA0FS0	*NCCS000*	*AQ0FS00*	*VMIP3S0*	*DA0MS0*	*NCMS000*	*SP*	*NCFS000*	*Fp*
	00694681-n	02957469-a	00800930-v		00766234-n		00962129-n	

Il·lustració 2*: Exemple d'anàlisi morfològica de l'oració "La justícia alemanya el delicte de rebel·lió."*

FONT: FreeLing Demo

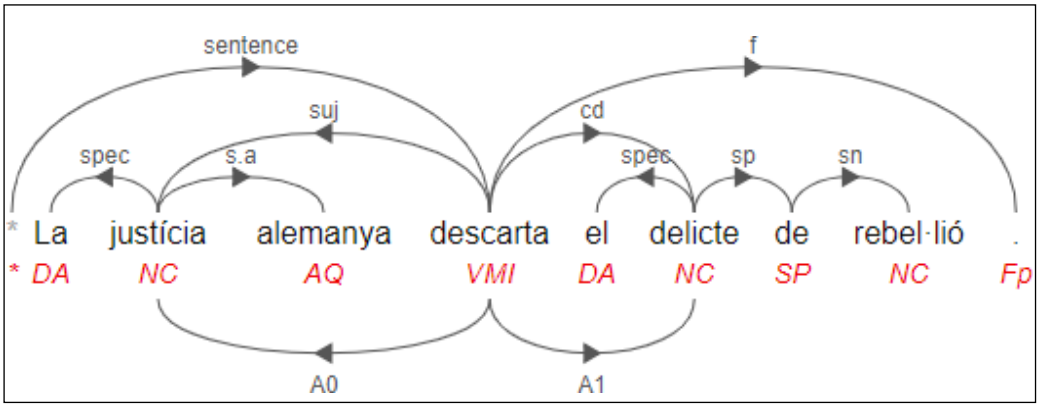
Aquest nivell d'anàlisi està directament relacionat amb el lèxic. El lèxic és el conjunt d'informació de cada paraula que el sistema utilitza per a realitzar el processament; hi inclou informació morfològica, la categoria gramatical, representació del significat i irregularitats sintàctiques.

2.3.2 Nivell sintàctic

Aquest nivell analitza les paraules d'una frase per tal de descobrir-ne l'estructura gramatical. La meva funció és aconseguir una representació de la frase que mostri les relacions de dependència entre les paraules. L'estructura sintàctica i la dependència contribueixen al seu significat. Per exemple, l'oració "Vaig veure un home amb uns prismàtics" varia el significat si s'analitza el sintagma "amb uns prismàtics" com a

complement del nom del SN “un home” o com a complement circumstancial d’instrument del verb “vaig veure”.

La ***Il·lustració 3*** mostra el resultat de l’anàlisi sintàctica, la qual atribueix les relacions de dependències entre les paraules i en descriu el tipus.



Il·lustració 3*: Exemple d’anàlisi sintàctica de l’oració “La justícia descarta el delicte de rebel·lió.”*

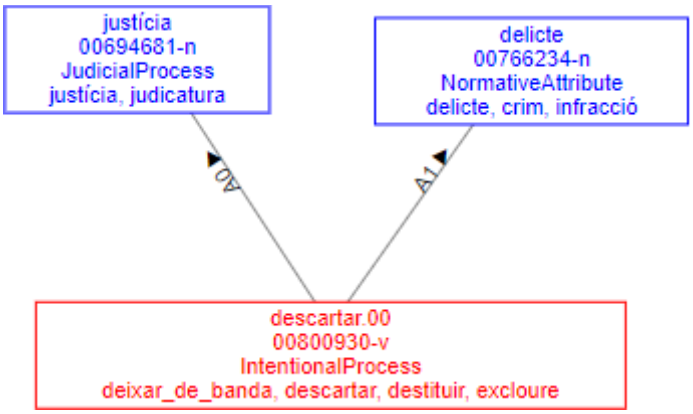
FONT: FreeLing Demo

2.3.3 Nivell semàntic

Tots els nivells d’anàlisi aporten significat a una oració o paraula, però el semàntic hi juga un paper fonamental. El tractament semàntic determina el significat d’una frase, perquè analitza els diferents nivells de significat de cada mot dins l’oració. També inclou la desambiguació de paraules polisèmiques: selecciona el sentit més adequat de la paraula.

El substantiu “banc”, per exemple, pot referir-se a un seient, una empresa financera o un conjunt de peixos. En aquest cas, si fos necessària la desambiguació i s’haguessin de tenir en compte altres elements de la frase, el nivell semàntic duria a terme aquesta funció, no el lèxic.

La ***Il·lustració 4*** mostra una representació de l’anàlisi semàntica, que indica el significat de cada mot i la manera amb què es combinen per formar el significat general de la frase. Aquesta anàlisi és independent de l’estructura sintàctica de l’oració, ja que només es refereix al significat de les paraules.



***Il·lustració 4**: Exemple de l'anàlisi semàntica de l'oració "La justícia alemanya descarta el delicte de rebel·lió." FONT:* FreeLing Demo

2.3.4 Nivell pragmàtic

Aquest nivell es refereix a l'ús del llenguatge de manera intencionada per atribuir un significat addicional al text. Per tant, es requereix molta informació perquè la comprensió d'intencions sigui viable. Considerem les dues oracions següents, per exemple:

"Els regidors van rebutjar els permisos als manifestants perquè (ells) tenien por de la possible violència".

"Els regidors van rebutjar els permisos als manifestants perquè (ells) defensaven la revolució".

Es requereix la resolució del terme "ells", que en la primera frase actua com a substitut de "Els regidors" i en la segona, de "els manifestants". Per dirimir aquest conflicte lingüístic es requereixen els coneixements del nivell pragmàtic.

2.4 Mètodes

Hi ha diferents tipus de mètodes del PLN en funció de la tècnica utilitzada per processar informació, els quals es poden classificar en quatre enfocaments: simbòlic, empíric o estadístic, connexionista i híbrid. Aquests quatre mètodes tracten la informació i els fenòmens lingüístics emprant diferents procediments depenent de la finalitat que es persegueix.

2.4.1 Enfocament simbòlic

Els enfocaments simbòlics es basen en regles del llenguatge generalment acceptades dins d'una llengua i en lèxics desenvolupats i enregistrats per lingüistes i que, posteriorment, s'agrupen en sistemes informàtics. Aquest enfocament realitza una anàlisi profunda dels fenòmens lingüístics i la seva representació mitjançant sistemes de representació del coneixement, elaborats i supervisats per experts.

Durant diverses dècades, s'ha utilitzat l'enfocament simbòlic en moltes àrees de recerca i en aplicacions com la resolució d'ambigüitats i l'extracció d'informació. L'aprenentatge basat en les regles i la Programació Lògica Inductiva[2] (ILP) són algunes de les tècniques més utilitzades.

2.4.2 Enfocament estadístic

Els enfocaments estadístics desenvolupen models generals i aproximats de fenòmens lingüístics basats en exemples proporcionats per diverses tècniques matemàtiques i corpus de text. En contrast amb els enfocaments simbòlics, els estadístics no afegeixen coneixements lingüístics, sinó que utilitzen dades observables i exemples comuns.

Aquest tipus d'enfocaments són més imprecisos, ja que es basen en aproximacions estadístiques a partir d'exemples. D'aquesta manera, si processen grans quantitats d'informació, es redueix la presència d'errors. A més a més, els enfocaments estadístics no requereixen la intervenció d'especialistes humans per especificar i desenvolupar models de fenòmens lingüístics. Aquest fet abarateix considerablement el cost econòmic i ha propiciat que les grans empreses hi mostrin interès i decideixin invertir-hi: Yahoo, Google, entre altres.

Un dels mètodes més comuns és el Model Ocult de Markov[3] (*Hidden Markov Model; HMM*), un model estadístic que, a partir de paràmetres observables, determina els paràmetres desconeguts. Aquest es pot utilitzar, per exemple, en aplicacions de

[2] L'ILP consisteix en l'aplicació del corpus de coneixement sobre lògica per al disseny de llenguatges de programació.

[3] Andrei Markov (1856-1922) fou un matemàtic rus conegut pels seus treballs en teoria de nombres i teoria de la probabilitat.

reconeixement de patrons i, en el nostre cas, en la determinació de fenòmens lingüístics.

2.4.3 Enfocament connexionista

La tècnica de l'enfocament connexionista utilitza xarxes neuronals per representar el coneixement lingüístic. Les xarxes neuronals (XNA) són un conjunt de tècniques d'aprenentatge automatitzat que simulen el processament cognitiu humà inspirat en el mecanisme de processament d'informació que empra el cervell. Les XNA estan formades per unitats -que creen un paral·lelisme amb les neurones- i per connexions entre elles mateixes.

Un dels principals avantatges d'aquest enfocament, comparant-lo amb els models chomskians, és l'absència de regles explícites. L'atribució de les categories gramaticals, per exemple, no s'ha de dur a terme de forma manual, sinó que la XNA en farà una aproximació de manera implícita. Aquest tret provoca que aquests sistemes posseeixin una alta tolerància de cara als errors. Si una oració és ambigua o té un verb mal conjugat, simplement es redueix l'eficiència. En les gramàtiques formals, en canvi, en aquests casos s'invalida totalment la frase o es crea una resposta sense coherència.

2.4.4 Enfocament híbrid

El PLN, inicialment, era de caràcter simbòlic i estava basat en regles i bases de coneixement. Més tard, va evolucionar i va popularitzar l'ús de tècniques estadístiques –l'ús de les màquines de vectors de suport, per exemple-. Com ja he explicat en cada cas, aquests dos enfocaments tenen, com qualsevol altre sistema, limitacions i fortaleses.

Amb la intenció de millorar els mecanismes del PLN i crear una metodologia més eficaç i potent que la dels anteriors sistemes, l'enfocament híbrid intenta combinar-los per així poder obtenir els punts més forts de cada un.

2.5 Aplicacions

El llenguatge natural és l'instrument que utilitzem els éssers humans per comunicar-nos. Avui en dia, gran part del coneixement humà es troba en format digital en diverses

col·leccions de dades. El PLN és important en aquest aspecte, ja que relaciona la comunicació humana amb aquest emmagatzematge de dades digitals i les modifica de diverses maneres en funció de l'objectiu que es persegueix. Hi ha moltes funcions que pot efectuar el PLN, algunes de les quals són descrites a continuació.

2.5.1 Recuperació i extracció d'informació

El procés de recuperació d'informació (RI) -*information retrieval* (*IR*)- consisteix en la transformació de dades no estructurades o semiestructurades, presents en documents o pàgines web, en representacions que segueixin uns models específics que concordin amb els criteris i propòsits de la cerca.

Un cop s'obtenen les dades, s'executa l'extracció d'informació (EI) –*information extracction* (*IE*)-. Aquest sistema es basa en el processament dels continguts anteriors en un format de base de dades, és a dir, d'una manera estructurada. En cada document es duu a terme aquest procés amb l'objectiu de trobar relacions de significat i contingut.

2.5.2 Mineria de dades textuals

La mineria de dades textuals (MDT) –*text data mining* (*TDM*)- és un procés que analitza i dedueix patrons i tendències que existeixen en les dades, que ja estan emmagatzemades en un format estructurat. Mitjançant anàlisis matemàtiques, també realitza un processament previ de normalització de la informació, que inclou la creació d'enllaços i taules.

2.5.3 Traducció automàtica

L'objectiu principal de la traducció automàtica (TA) –*machine translation* (*MT*)- és traduir una informació específica a un altre idioma sense canviar-ne el significat. És una de les aplicacions més antigues del PLN i, malgrat els evidents avanços tecnològics en aquest àmbit, encara hi ha molts objectius per complir. Per tant, queda un llarg recorregut per continuar la recerca, especialment en llengües amb l'ordre de les paraules invers o una morfologia molt complexa.

Sovint apareixen problemes de significat entre expressions, però mitjançant un enfocament estadístic de diverses repeticions de casos, s'aconsegueixen avaluacions i resultats precisos.

2.5.4 Sistemes de cerca de respostes

El sistema de cerca de respostes –*question-answering* (*QA*)- proporciona una resposta aparentment natural – "Siri", d'Apple-. D'una banda, la base d'aquests sistemes és la cerca i la gran quantitat de contingut per obtenir una resposta adequada i, de l'altra, utilitzen mètodes estadístics per entendre les preguntes de l'usuari i determinar-ne el tipus de resposta.

2.5.5 Generació de resums automàtics

El nivell més alt del PLN, concretament el de discurs[4], és capaç d'implementar una reducció d'un text llarg a una representació abreujada del document original.

La generació de resums automàtics (*summarization*) s'efectua a escala de textos particulars: s'identifiquen els paràgrafs i es recullen els termes més importants que defineixen el significat de l'article original. També es poden resumir col·leccions de documents, on s'executa un procés d'agrupació de tòpics i d'identificació de similituds i diferències de la informació que contenen, valorada des d'un punt de vista semàntic.

2.5.6 Sistemes de diàleg

Els sistemes de diàleg (*dialogue systems*) es consideren una de les propostes amb més futur segons els proveïdors d'aplicacions. Aquests tenen la capacitat de comprendre i mantenir un diàleg amb una aplicació específicament definida -com per exemple, una nevera o un sistema de so domèstic-. Actualment, s'utilitzen els nivells fonètic i lèxic del llenguatge, però tots els nivells ofereixen prou potencial per establir sistemes de diàleg.

[4] El nivell de discurs es centra en les propietats del text en el seu conjunt que transmet significat fent connexions entre les oracions del component.

Per exemple, els articles de diaris es poden desconstruir en components del discurs com ara: títol, història principal, esdeveniments anteriors, avaluació i expectatives.

3. Part pràctica: programa analitzador de notícies

Amb l'objectiu d'exemplificar la part teòrica i demostrar una de les aplicacions del PLN, he utilitzat un programa que aplica tècniques d'extracció i anàlisi d'informació a diferents articles i notícies de diaris. Trobareu la versió completa d'aquest *software* a l'apartat d'annexos.

3.1. Descripció i objectiu

Mitjançant l'ús de la plataforma *Anaconda,*[5] he utilitzat un programa implementat en el llenguatge de programació *Python* que es basa en l'anàlisi d'informació per extreure resultats que derivin en conclusions.

He recollit un total de 120 notícies de catorze diaris diferents en tres idiomes (català, castellà i anglès) relacionades amb la figura del president Puigdemont i diversos fets concrets i generals del mateix tema.

El programa informàtic, mitjançant servidors externs com el *FreeLing* o l'*NLTK* –explicats més endavant- processa tota la informació extreta de les notícies del diari. Posteriorment elabora una taula de dades amb les paraules i estructures gramaticals més associades a cada diari.

D'aquesta manera, analitzant els resultats es poden extreure conclusions sobre la ideologia de cada diari, valorar-ne l'objectivitat i/o imparcialitat de la premsa a l'hora d'informar sobre un fet concret o tractar un tema d'àmbit polític, com és en aquest cas.

3.2. El llenguatge de programació *Phyton*

Python és un llenguatge de programació de propòsit general que s'ha popularitzat molt durant l'última dècada. És interactiu i empra una sintaxi clara i visual, de maneraque

[5] Anaconda és una distribució *freemium* de codi obert dels llenguatges de programació *Python* i R per al processament de dades de gran escala, analítica predictiva i computació científica, que tracta de simplificar la gestió i desplegament de paquets.

es poden crear programes de forma organitzada i senzilla. Moltes de les grans empreses com Google, Yahoo o la NASA l'utilitzen i es preveu que se segueixi expandint en els anys vinents.

3.3. Execució del programa

En primer lloc, el programa (a la ***Il·lustració 5*** en presento un fragment) cerca la notícia, un text sense format en codificació UTF-8[6] en un directori especificat anteriorment, com per exemple: *noticies-ca-totes*.

A continuació, se segmenten les frases de cada notícia i s'executa la tokenització: extracció de tots els *tokens* del text –cadenes de caràcters que tenen un valor o són unitats independents, com qualsevol paraula, un número o un punt-.

Un cop extrets tots el *tokens* es realitza el *tagging* de cada notícia, que equival a atribuir etiquetes morfosintàctiques a les paraules o *tokens*. Per dur a terme aquestes funcions de segmentació de frases, tokenització i etiquetatge morfosintàctic s'utilitza una plataforma externa anomenada *FreeLing*.

El següent pas és la intervenció del *parser* (analitzador sintàctic), una eina que prové de la plataforma *NLTK*. Aquest analitzador aplica a cada una de les frases etiquetades morfosintàcticament les regles i les gramàtiques del *chunker*[7] (ja definides anteriorment i que identifiquen paraules i sintagmes).

Tot seguit, s'analitzen diversos factors, com el nombre d'aparicions totals de cada paraula o sintagma, el nombre d'aparicions totals de cada paraula o sintagma en cada diari, el nombre de paraules en les notícies de cada diari i el nombre de paraules totals.

Finalment, amb aquestes dades es calculen les mètriques de rellevància. Els resultats es guarden en un directori de sortida: *noticies-ca-totes-out*, en aquest cas.

```
def analitza(lang, dir_in, minaparicions, result):
    aparicionstotal = defaultdict(int)   # num aparicions totals de cada paraula
    aparicionsdiari = {}                 # num aparicions de cada paraula en cada diari
    longituddiari = defaultdict(int)     # num de paraules en les noticies d'un diari
    longitudtotal = 0.0                  # num de paraules en totes les noticies
```

[6] UTF-8 (8-bit *Unicode Transformation Format*) és un format de codificació de caràcters.

[7] *Chunker*: analitzador per fragments.

```
dir_out = dir_in+"-out"    # directori on deixar les paraules extretes
```

```
for f in os.listdir(dir_in) :
    print ("tractant fitxer"+dir_in+"/"+f)
    # especifica els fitxers d'entrada i de sortida
    noticia = open(dir_in+"/"+f, encoding='utf-8-sig').read()
    sortida = open(dir_out+"/"+f, 'w')



# nom del diari
   nomdiari = f.split("_")[0]
  if   nomdiari   not   in   aparicionsdiari   :   aparicionsdiari[nomdiari]   =
       defaultdict(int)

    # fer el tagging de la noticia
    frases_etiquetades = freeling(lang, noticia)

    # gramàtica i parser (analitzador sintàctic)
    parser = nltk.RegexpParser(gramatica[lang])


    # per a cadascuna de les frases etiquetades  aplicar les regles del parser
    for sent in frases_etiquetades:
        tree = parser.parse(sent)


        for subtree in tree.subtrees():
            # copia subarbres reconeguts per la gramatica al fitxer de sortida,
            # i incrementa els comptadors.
            if (subtree.label() != 'S') :
                print(subtree,  file=sortida)
                s = ""
                for x in subtree.leaves() : s += x[0].lower()+"_"
                s = s[:-1]
                aparicionsdiari[nomdiari][s] += 1
                aparicionstotal[s] += 1
                longituddiari[nomdiari] += 1
                longitudtotal += 1.0

    sortida.close()
```

Il·lustració 5*: Fragment del programa analitzador de notícies elaborat amb* Python*.*

FONT: Elaboració pròpia

3.4. Plataformes externes

3.4.1. FreeLing

FreeLing[8] és una biblioteca *online* que proporciona funcionalitats d'anàlisi lingüística (separació de frases, anàlisi i desambiguació morfològica, detecció d'entitats, desambiguació del sentit de la paraula, etiquetatge de rols semàntics, etc.) per a

[8] http://nlp.lsi.upc.edu/freeling/index.php/node/1

diverses llengües (anglès, espanyol, portuguès, italià, alemany, rus, català, gallec, croat i eslovè, entre d'altres).

Com ja he explicat anteriorment, al meu programa utilitzo aquesta plataforma per segmentar les frases i identificar i etiquetar els *tokens* o les paraules segons la seva categoria gramatical.

El	periple	europeu	de	Puigdemont	acaba	després	de	cent	trenta-dos	dies	.
el	periple	europeu	de	puigdemont	acabar	després	de	ce_nit	trenta-dos	dia	.
DA0MS0	*NCMS000*	*AQ0MS00*	*SP*	*NP00000*	*VMIP3S0*	*RG*	*SP*	*NCFS000*	*RG*	*NCMP000*	*Fp*

Il·lustració 6: *Anàlisi morfosintàctica de l'oració: "El periple europeu de Puigdemont acaba després de cent trenta-dos dies." mitjançant* FreeLing*. FONT:* FreeLing demo

Com es pot observar a la ***Il·lustració 6***, *FreeLing* codifica la informació morfològica en etiquetes *PoS,*[9] les quals es basen en les propostes d'*EAGLES.*[10]

Les etiquetes de PoS d'*EAGLES* consisteixen en marques de longitud variable en les quals cada caràcter correspon a una característica morfològica. El primer caràcter de l'etiqueta sempre és la categoria que determina la llargada i la interpretació de cada caràcter de l'etiqueta.

Els atributs que no són aplicables o no es concreten en una paraula específica es defineixen amb un 0. Per tant, l'etiqueta "NCMS000" que *FreeLing* atorga al mot "periple" (***Il·lustració 6***) indica que es tracta d'un nom comú masculí singular, però les altres tres posicions següents no són específiques.

[9] *Part of Speech* (*PoS*)*:* categoria morfològica.

[10] *EAGLES* (*Expert Advisory Groups on Language Engineering Standards*) és una iniciativa recolzada per la Unió Europea que té l'objectiu de normalitzar la pràctica del PLN.

Part of Speech: noun

Position	Atribute	Values
0	category	**N**:*noun*
1	type	**C**:*common*; **P**:*proper*
2	gen	**F**:*feminine*; **M**:*masculine*; **C**:*common*
3	num	**S**:*singular*; **P**:*plural*; **N**:*invariable*
4	neclass	**S**:*person*; **G**:*location*; **O**:*organization*; **V**:*other*
5	nesubclass	*Not used*
6	degree	**V**:*evaluative*

***Taula 1**: Definició de l'etiquetes de la categoria "NOM" per al català. Així és poden crear etiquetes de* PoS *com per exemple "NCMS000" (Nom, Comú, Masculí, Singular). FONT:* FreeLing Tagset Description

Les etiquetes solen variar en funció de l'idioma, ja que les característiques morfològiques són diferents a les catorze llengües que engloba *FreeLing*.

3.4.2. NLTK

NLTK[11] *(Natural Language Toolkit)* és una plataforma *online* que permet elaborar programes amb llenguatge de programació *Python* destinat a aplicacions de PLN. Inclou biblioteques i programes de *tokenització*, *taggers* (etiquetadors morfològics), *parsers* (analitzadors sintàctics), demostracions gràfiques i conjunts de dades de mostra.

El meu programa utilitza el *parser* d'aquesta plataforma per elaborar regles sintàctiques i gramàtiques per formar sintagmes.

3.5. Gramàtiques del *chunker*

L'objectiu del programa és analitzar els textos i detectar les paraules o els sintagmes que apareguin més cops o que s'utilitzin amb més freqüència. Les paraules queden

[11] https://www.nltk.org/

definides amb el *tokenitzador* i l'analitzador morfosintàctic, però, els sintagmes s'han de definir manualment mitjançant expressions regulars per així formar les regles de les gramàtiques del *chunker*.

Per al programa analitzador de notícies, he definit gramàtiques de *chunks* en tres idiomes (castellà, català i anglès). Tot i això, el programa finalment només utilitzarà la gramàtica del català, tal com es mostra en la ***Il·lustració 7***, ja que he traduït totes les notícies a aquest idioma per facilitar-ne la comparació i obtenir millors resultats.

```
gramàtica["ca"] = r"""
    N: {<NC.*|NP.*>}
    SA: {<RG>*<AQ.*|VMP.*>}
    SV: {<VM.*><D.*><N>}
    SN: {<N><SP.*><D.*>?<N>}
        {<D.*>?<SA>+<N>+}
        {<D.*>?<N>+<SA>+<N>?}
```

Il·lustració 7*: Gramàtica del chunker extreta del meu programa i utilitzada per definir sintagmes en català. FONT: Elaboració pròpia*

3.5.1. Expressions regulars

Una vegada definides les etiquetes morfològiques de cada paraula, les hem d'ajuntar per formar sintagmes. Per executar aquesta fusió, s'utilitzen les anomenades "expressions regulars".

Una expressió regular és una descripció compacta d'un conjunt de caràcters, paraules o patrons de text, que es formen gràcies a l'ús d'unes regles sintàctiques.

Les expressions regulars poden consistir en:

1. Seqüències de caràcters literals, per exemple 'NCFS000'
2. Disjuncions de caràcters que s'expressa amb "|" com 'NCF.*|NCM.*', on el punt (".") fa referència a qualsevol caràcter.
3. Comptadors que es poden referir al caràcter o al grup precedent:
    a. *: apareix cap cop o més
    b. +: apareix un cop o més
    c. ?: és opcional

3.5.2. Regles de la gramàtica desenvolupades

Aquestes són les regles que he utilitzat per identificar els termes –paraules i sintagmes- en les notícies recollides.

```
N: {<NC.*|NP.*>}
```

***Regla 1**:* Defineix la partícula "N" (nom), que pot estar formada per un nom comú de qualsevol gènere o nombre (<NC.*>) o un nom propi que segueix el mateix patró (<NP.*>). La disjunció dels dos caràcters es fa mitjançant el símbol '|'.

```
SA: {<RG>*<AQ.*|VMP.*>}
```

***Regla 2**:* Defineix un sintagma adjectival compost per un adverbi que pot no aparèixer o fer-ho més d'una vegada (<RG>*) i un adjectiu qualificatiu o bé un verb en forma de participi (<AQ.*|VMP.*>); els dos en qualsevol gènere i nombre.

```
SV: {<VM.*><D.*><N>}
```

***Regla 3**:* Defineix un sintagma verbal compost per un verb principal (<VM.*>), un determinant –article, demostratiu, indefinit, possessiu, relatiu, interrogatiu o exclamatiu- en qualsevol gènere i nombre (<D.*>) i un nom (partícula <N> definida anteriorment.

```
SN: {<N><SP.*><D.*>?<N>}
    {<D.*>?<SA>+<N>+}
    {<D.*>?<N>+<SA>+<N>?}
```

***Regla 4**:* Defineix tres tipus de sintagmes nominals diferents. El primer està compost per un nom, un sintagma preposicional (<SP.*>), un determinant de qualsevol classe, gènere i nombre que és opcional -per tant es pot aplicar la regla sense l'obligació que hi aparegui un determinant (<D.*>?)- i un altre nom.

El segon sintagma està format per un determinant opcional de qualsevol classe, gènere i nombre, un o més sintagmes adjectivals que han estat definits anteriorment i un nom que pot aparèixer un cop o més (<N>+).

Finalment, el tercer sintagma el formen un determinant opcional com els dos anteriors, un nom i un SA (sintagma adjectival) -que poden aparèixer un cop o més- i un altre nom també opcional (<N>?).

D'aquesta manera, el programa utilitza les gramàtiques creades per extreure sintagmes com el de la

Il·lustració 8.

```
(SN (N president/NCMS000) de/SP la/DA0FS0 (N Generalitat/NP00000)
(SN l'/DA0CS0 (N Estat/NP00000) (SA espanyol/AQ0MS00))
(SN (N investidura/NCFS000) a/SP (N distància/NCFS000))
(SN (N eines/NCFP000) (SA polítiques/AQ0FP00) (SA diferents/AQ0CP00))
```

Il·lustració 8*: Exemples de frases en notícies dels diaris* ARA *i* El Nacional *de la gramàtica del* chunk *definida a la* ***Regla 4***

```
(SV denegar/VMN0000 l'/DA0CS0 (N escorta/NCCS000))
(SV usar/VMN0000 els/DA0MP0 (N símbols/NCMP000))
(SV defensar/VMN0000 la/DA0FS0 (N democràcia/NCFS000))
```

Il·lustració 9*: Exemples de frases de notícies de* La Razón *i* El Nacional *de la gramàtica del* chunk *definida a la* ***Regla 2***

3.6. Diaris i notícies

La informació recollida i que ha analitzat el programa que he codificat, tal com he especificat anteriorment, està formada per 120 notícies de catorze mitjans de comunicació diferents. D'aquests rotatius, quatre són editats en català (*Ara, El Periódico, El Punt Avui i Vilaweb*), vuit en castellà (*ABC, El Mundo, El Nacional, El País, La Razón, La Vanguardia, Libertad Digital i OK Diario*) i dos en anglès (*Independent i The Guardian*).

Tot i ser diaris editats en tres idiomes diferents, a l'hora d'analitzar-los i comparar-los han d'estar tots en el mateix idioma perquè el programa funcioni i n'obtingui resultats. Per tant, he traduït les 75 notícies al castellà, i les 10 en anglès, al català mitjançant l'eina *online Google Translator.*[12]

Pel que fa a la recerca de notícies, m'he basat en la figura del president Carles Puigdemont i en quatre esdeveniments concrets relacionats amb la seva figura: la

[12] https://translate.google.es

compareixença de Puigdemont als jutjats belgues (de l'1 al 7 de novembre), l'entrada de Puigdemont a la presó alemanya (del 24 al 29 de març), el descart de delicte de malversació per part de la justícia alemanya (del 3 al 6 d'abril) i la sol·licitud d'extradició de Puigdemont de la fiscalia alemanya (principis de juny).

Aquestes notícies, que corresponen al 50% (60 textos) de les totals, permeten veure com diferents diaris anuncien un mateix esdeveniment des de diferents punts de vista amb graus d'objectivitat diversos.

La resta de notícies no se centren en fets concrets, sinó que tracten la figura de Carles Puigdemont des de l'àmbit polític general. D'aquesta manera, es potencia i s'augmenta el vocabulari propi de cada diari i s'aconsegueixen resultats més precisos.

La xifra de mots totals analitzats a les 120 notícies és de 31.162. La ***Taula 2*** mostra el nombre de paraules de cada un dels diaris utilitzats. Aquests valors -les paraules totals i les pròpies de cada diari- seran útils més endavant per calcular el percentatge de rellevància dels termes.

DIARI	NÚMERO PARAULES DIARI	DIARI	NÚMERO PARAULES DIARI
ABC	2.540	*INDEPENDENT*	713
ARA	3.245	*LA RAZÓN*	2.222
EL MUNDO	2.497	*LA VANGUARDIA*	2.732
EL NACIONAL	1.264	*LIBERTAD DIGITAL*	2.610
EL PAÍS	4.282	*OKDIARIO*	2.326
EL PERIÓDICO	1.773	*THE GUARDIAN*	1.596
EL PUNT AVUI	1.898	*VILAWEB*	1.464

Taula 2: *S'indica el nombre de paraules analitzades de cada un dels catorze diaris utilitzats.*

3.7. Resultats i interpretació

Després d'analitzar els textos i de realitzar totes les accions anteriors, el programa n'extreu uns resultats anomenats *output*. Aquests resultats es mostren en forma de dues mesures: *Tf-idf* i *IM*.

El programa ha estat executat de tres maneres diferents per així obtenir tres nivells de resultats. El primer, agrupant només les notícies d'un mateix esdeveniment concret. El

segon nivell, ajuntant tota la informació (60 textos) dels quatre fets descrits anteriorment. Finalment, un tercer nivell que engloba tot el conjunt de les notícies.

D'aquesta manera, he pogut observar que amb més quantitat d'informació, els resultats guanyen precisió i permeten una millor interpretació de la ideologia o del punt de vista que té cada diari sobre el tema.

3.7.1. Tf-idf

Tf-idf (de l'anglès *Term frequency – inverse document frquency*) és una mesura usada en la recuperació de la informació per determinar la rellevància d'un terme en una consulta (és a dir, la paraula buscada al cercador) respecte un document que forma part de la resposta de la cerca.

En general, es pot definir com una mesura de rellevància d'un terme sobre un subconjunt de textos que formen part d'una col·lecció (cada subconjunt pot ser un sol document o un grup de documents). En el meu cas vull calcular si un terme és rellevant per a un diari; per tant tindré en compte la col·lecció formada per totes les notícies recollides i els subconjunts formats per les notícies de cada diari. La mesura es basa en dos components: *Tf* (*Term frequency*) i l'*idf* (*Inverse Document Frequency*).

El *Tf* mesura si el terme és freqüent al subconjunt considerat (un diari en el meu cas). En principi, com més alta és la freqüència d'un terme en un diari, més incrementa la seva rellevància.

Tanmateix, hi ha termes molt freqüents a la llengua, com per exemple preposicions, articles o verbs comuns ("de", "la", "fer", "haver", "ser", "cosa", etc.). Per tant, si un terme apareix a molts diaris, la seva rellevància disminueix i es converteix en un terme general i comú.

D'aquesta manera, l'*idf* mesura la quantitat de diaris diferents en els quals apareix el terme, comptant el percentatge de rotatius totals.

La rellevància ha de decréixer quan el percentatge de diaris en els quals apareix el terme augmenta; s'usa la inversa d'aquest percentatge (és a dir, l'*idf*).

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

***Il·lustració 10**: 'tf' són el nombre d'aparicions del terme 'x' en el diari 'y', 'N' és el nombre de diaris diferents de la col·lecció i 'df' són el nombre de diaris diferents on apareix el terme 'x'.*

FONT: FiloTechnologia

3.7.2. Informació Mútua

La Informació Mútua (IM) és una mesura que s'utilitza per calcular el grau de dependència entre dues variables aleatòries. En el meu cas, l'aparició d'un terme concret en una notícia em pot donar informació sobre el diari en què apareix.

D'aquesta manera, per calcular la informació que em dóna un terme i un diari en particular es tenen en compte tres variables diferents:

- La probabilitat que el terme *t* aparegui al diari *d*:

  p(d,t) = núm. aparicions terme t en el diari d / núm. total de termes apareguts en el diari d

- La probabilitat que el terme *t* aparegui a la col·lecció:

  p(t) = núm. aparicions terme t a la col·lecció / núm. total de termes apareguts en la col·lecció

- La probabilitat que una aparició d'un terme qualsevol de la col·lecció sigui del diari *d*:

  p(d) = núm. total de termes apareguts en el diari d / núm. total de termes apareguts en la col·lecció

Si un cop calculat el coeficient entre aquests nombres s'obté un valor més gran que 1 significa que *t* i *d* coincideixen en una freqüència més elevada de la mitjana; si el valor és inferior a 1, la freqüència observada és menor a la que s'esperava.

3.7.3. Taules

Per tal d'extreure conclusions dels resultats del programa, he elaborat catorze taules – una per cada rotatiu- amb els valors de la mesura *IM*, ja que els valors negatius permeten aprofundir en l'extracció d'informació (el conjunt sencer de les catorze taules es troba als annexos). A la ***Taula 3*** i la ***Taula 4*** s'observen els resultats extrets de les notícies dels diaris *Ara* i *OK Diario*.

Les taules estan formades per dues parts: una amb els valors positius, on s'indiquen les paraules més rellevants per al diari; i una amb els valors negatius, en la qual apareixen els termes menys utilitzats o amb menor importància de cada font.

He remarcat de color verd tots els mots que estan relacionats amb el tema en qüestió, els termes amb major rellevància i subjectivitat i/o altres paraules que poden influir a l'hora d'obtenir conclusions.

ARA							
VALOR POSITIU				**VALOR NEGATIU**			
APD	**APT**	**IM**	**PARAULA**	**APD**	**APT**	**IM**	**PARAULA**
3	**3**	**3.26**	**consellers_a_brussel·les**	**1**	**23**	**-1.26**	**preventiva**
3	**3**	**3.26**	**ordre_de_recerca**	**1**	**23**	**-1.26**	**independent**
4	**4**	**3.26**	**captura_internacional**	1	24	-1.32	acte
4	4	3.26	interlocutòries	1	25	-1.38	grup
4	4	3.26	els_pròxims_dies	**1**	**25**	**-1.38**	**advocacia_de_l'_estat**
3	**3**	**3.26**	**il·lícita**	**1**	**25**	**-1.38**	**l'_estat_espanyol**
4	4	3.26	aportades	1	26	-1.44	quim_torra
3	3	3.26	döpper	1	27	-1.49	premsa
3	**3**	**3.26**	**causa_contra_el_procés**	**1**	**27**	**-1.49**	**querella**
4	**4**	**3.26**	**poder_judicial**	4	114	-1.57	català
3	**3**	**3.26**	**exconsellers_a_l'_exili**	1	29	-1.59	possible
4	5	2.94	pròxims	**1**	**31**	**-1.69**	**tribunals**
3	**4**	**2.85**	**part_de_el_govern**	**1**	**31**	**-1.69**	**ciutadans**
6	**8**	**2.85**	**audiència_territorial**	**1**	**32**	**-1.74**	**la_justícia_alemanya**
6	8	2.85	ara				
3	**4**	**2.85**	**proves**				
3	**4**	**2.85**	**transversal**				
5	**7**	**2.78**	**costos**				
5	7	2.78	pròximes				
2	**3**	**2.68**	**el_ministeri_públic**				
2	3	2.68	interpol				
2	**3**	**2.68**	**comú_acord**				
2	**3**	**2.68**	**alçament**				
2	3	2.68	organitzat				
2	3	2.68	associacions				
2	**3**	**2.68**	**la_policia_federal_belga**				
4	6	2.68	anunci				
2	**3**	**2.68**	**cas_de_puigdemont**				
2	**3**	**2.68**	**el_tribunal_regional**				
2	**3**	**2.68**	**demanda_de_puigdemont**				

***Taula 3**:Taula elaborada amb els resultats de les notícies del diari* Ara *amb els valors de la mesura IM.*

OK DIARIO							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	3	3.74	dolors_bassa	**1**	**27**	**-1.01**	**querella**
3	**3**	**3.74**	**exconseller**	**5**	**147**	**-1.13**	**espanyol**
3	**3**	**3.74**	**fugida_de_el_colpista**	1	31	-1.21	general
8	**8**	**3.74**	**colpista**	**1**	**32**	**-1.26**	**la_justícia_alemanya**
3	**3**	**3.74**	**falsificat**	**1**	**32**	**-1.26**	**civil**
3	3	3.74	oral	4	135	-1.33	extradició
4	**5**	**3.42**	**seguiment**	**4**	**139**	**-1.38**	**tribunal**
2	**3**	**3.16**	**exconsellers_fugats**	1	35	-1.39	portaveu
2	**3**	**3.16**	**desobediència_a_l'_autoritat**	1	36	-1.43	càrrec
4	6	3.16	oede	1	36	-1.43	passat
2	3	3.16	meritxell_borràs	**1**	**37**	**-1.47**	**eleccions**
2	3	3.16	treball	1	38	-1.50	nou
2	3	3.16	santi_vila	1	39	-1.54	membres
2	**3**	**3.16**	**reiterades**	**5**	**198**	**-1.56**	**president**
2	**3**	**3.16**	**previsible**	**1**	**40**	**-1.58**	**independentisme**
2	3	3.16	viatjat	**1**	**42**	**-1.65**	**presos**
2	3	3.16	feina	**1**	**43**	**-1.68**	**lliurament**
2	**3**	**3.16**	**impunitat**	1	45	-1.75	erc
2	**3**	**3.16**	**cooperació_policial**	**1**	**72**	**-2.43**	**violència**
2	3	3.16	currículum				
2	3	3.16	organismes				
2	**3**	**3.16**	**conselleria_d'_interior**				
2	**3**	**3.16**	**serveis_de_seguretat**				
7	11	3.09	document				
3	**5**	**3.01**	**incondicional**				
6	**11**	**2.87**	**colpistes**				

Taula 4**: Taula elaborada amb resultats de les notícies de l'*OK Diario *amb els valors de la mesura IM.

3.7.4. Interpretació de les taules

La ideologia de cada mitjà de comunicació es pot identificar, per exemple, analitzant les etiquetes i els adjectius que es refereixen a Carles Puigdemont, la figura referencial d'aquest estudi. Un cas molt clar a simple vista és el de les paraules relacionades amb el terme 'fuga' o 'alçament'. L'*OK Diario,* l'*ABC*, *Libertad Digital* i *Independent* utilitzen els mots "fugitiu", "cop d'estat", "escapada", "colpista", "rebel" o "secessionista" per dirigir-se al polític català i les seves actuacions.

L'altra cara de la moneda pertany a diaris com *El Nacional*, l'*Ara*, o *The Guardian.* Aquí les fonts utilitzen termes com 'president a l'exili' o 'exiliat'.

Un altre factor clau és la manera d'enfocar el càrrec de Puigdemont al govern català. Comprovem com se'l tracta de president o d'expresident. Un exemple es pot apreciar gràcies a la comparació de *La Vanguardia* i l'*ABC*. Mentre el primer fa ús del sintagma 'el president cessat', el segon utilitza 'l'expresident cessat'. A més a més, els termes 'president' i 'consellers' apareixen a la taula de valors positius en diaris com *El Nacional*, *La Vanguardia*, *El Punt Avui*, i l'*Ara* i a la de valors negatius en *El Mundo*, *Independent*, *OK Diario*, l'*ABC*. Així, els termes oposats 'expresident' i 'exconsellers' s'intercanvien a les taules de les fonts anteriors.

També és interessant observar l'abundància de termes relacionats amb el concepte 'presó', 'arrest', 'detenció' i 'desobediència' en publicacions de *La Razón*, l'*OK Diario*, *Libertad Digital*, *El Mundo* i *Independent*. Aquests termes, en canvi, sovintegen en la taula de valors negatius en fonts com *El Punt Avui*, *Vilaweb* o *El Nacional*. En aquests diaris apareixen amb més freqüència les paraules "llibertat" i/o "alliberament."

Cal destacar el contrast entre els mots 'fractura', 'divisió' i 'caiguda' que inclouen els mitjans com *El Periódico*, *El Mundo* o *The Guardian* i les paraules emprades per articles de l'*Ara* i que apareixen amb un valor negatiu en *El Mundo* i *El País*: 'acord' 'acord comú' i 'suport'.

Finalment, els termes relacionats amb 'dret' o 'drets fonamentals' apareixen amb un valor negatiu inferior a -1,40 en diaris com *La Razón*, *Libertad Digital* i *El País*; i amb un valor superior a 4,04 en *El Nacional*.

4. Entrevistes

4.1. Entrevista a Lluís Padró Cirera

Lluís Padró és professor i membre del Grup de Tractament del Llenguatge Natural TALP[13] al Departament de Programari de la Universitat Politècnica de Catalunya (UPC). Imparteix docència en aquesta universitat des de 1991, és doctor en Ciències de la Computació des de 1998 i ha dirigit més de trenta projectes de graduació en Enginyeria de Programari. La seva experiència docent engloba un ampli conjunt de matèries en cursos de Grau de Computació de la UPC per a estudiants de Programari, Telecomunicacions i Enginyeria civil.

Il·lustració 11: *Lluís Padró*

La seva àrea de recerca se centra bàsicament en la Intel·ligència Artificial, concretament en la construcció d'analitzadors de llenguatges del PLN. Té desenes de publicacions en revistes nacionals i internacionals (*Machine Learning*, *Computers & the Humanities*) i ha participat en nombroses conferències.

També és el principal desenvolupador i administrador de *FreeLing,* un conjunt de programari lliure que proporciona funcions d'anàlisi lingüística per a textos en diversos idiomes -jo he fet ús d'aquest analitzador en el meu programa i l'he utilitzat com a exemple per explicar els nivells d'anàlisi del PLN del marc teòric-.

La seva extensa trajectòria i experiència en el món del PLN i els seus coneixements en l'àmbit de la IA han permès l'elaboració d'una entrevista d'alt nivell.

La meva intenció és conèixer i analitzar el seu parer –un enfocament informàtic- sobre diferents aspectes de la Lingüística Computacional.

Les diverses trobades dutes a terme durant els mesos d'agost i setembre i a l'entrevista realitzada el dia 31 d'agost –transcrita a l'apartat d'annexos-, m'han mostrat la seva visió sobre aquest àmbit, que es resumeix a continuació.

[13] *Center for Language and Speech Technologies and Applications*.

El professor Padró destaca la importància de la part informàtica dins la Lingüística Computacional, ja que, segons ell, l'aspecte lingüístic pot ser substituït per mètodes estadístics, els quals tenen aquests coneixements en la codificació de les dades. D'aquesta manera, associa aquesta disciplina a un àmbit de "ciències", contràriament amb la visió general.

També fa una comparació entre l'enginyeria (informàtica) i la ciència (lingüística) amb diversos fets històrics per exemplificar la seva teoria: "La catapulta es va inventar abans de conèixer la llei de la gravetat i la fórmula del tir parabòlic. També, la combinació química de compostos (per crear sabó, per exemple) es feia molt abans de conèixer la teoria atòmica sobre com es combinen les molècules."

Finalment, l'informàtic fa una previsió de l'evolució de la Intel·ligència Artificial que, des del seu punt de vista, estarà liderada per les xarxes neuronals i, per tant, per l'enfocament connexionista. Tot i aquesta afirmació, no creu que aquesta revolució del PNL es dugui a terme en un futur imminent, ja que encara hi ha moltes qüestions que no tenen resposta.

4.2. Entrevista a Montserrat Marimon Felipe

Il·lustració 12: Montserrat Marimon

Montserrat Marimon és investigadora sènior del Centre de Supercomputació de Barcelona (BSC), especialitzada en la branca de la Mineria de Text. També és docent a la Universitat Pompeu Fabra (UPF) i treballa al Departament de Traducció i Ciències del Llenguatge.

El seu camp de recerca es basa, principalment, en la Intel·ligència Artificial i el desenvolupament del PLN enfocat a objectius d'àmbit lingüístic, com per exemple la creació de corpus sintàctics i semàntics. També ha escrit més d'una vintena d'articles sobre Lingüística Computacional.

Tal com podem llegir a l'entrevista efectuada el dia 22 d'octubre i que es reprodueix a l'apartat d'annexos, la lingüista destaca el seu aprenentatge autodidàctic en el camp de la programació informàtica a partir d'una màquina Unix[14], de sistemes de

[14] Unix és un sistema operatiu portable, multifuncional i multiusuari creat el 1969 per AT&T.

gramàtiques i, més endavant, de cursos en línia. Tot i això, no es penedeix de no haver fet un grau informàtic, ja que opina que la lingüística és la base de la Lingüística Computacional.

Marimon remarca l'increment de la quantitat d'aplicacions ofertes pel PLN en els darrers anys i l'ús que en fan les grans empreses. D'aquestes, en destaca els traductors automàtics, els *chatbots* d'atenció al client i els gestors de documents. No obstant això, reconeix que les companyies es basen principalment en la part informàtica, encara que cada cop incorporen més lingüistes.

També afirma, confirmant la teoria de l'enginyer Padró, que les xarxes neuronals –mitjançant el *Deep Learning*[15]- es convertiran en el referent del PNL en els pròxims anys. Seguint aquesta idea, preveu que el paper dels lingüistes es reduirà simplement a estructurar i codificar les dades necessàries per entrenar les xarxes neuronals i, paral·lelament, es mantindrà la investigació de models en l'àmbit acadèmic.

Per concloure, assegura que la lingüística no és una disciplina de lletres pures, sinó que inclou molts conceptes que s'associen amb un àmbit científic, com la seva base formal i els models teòrics matemàtics que utilitza. A més a més, l'aparició i el creixement d'Internet al llarg de l'última dècada –fet que ofereix una quantitat enorme de dades de la llengua- ha comportat la consolidació de la part experimental de la disciplina.

[15] *Deep Learning* és una tècnica d'extracció i transformació de noves característiques del processament de la informació, les quals poden ser de forma supervisada o no.

5. Conclusions

Arribats en aquest punt del treball, puc dir que he assolit els objectius, específics i generals, que em vaig proposar inicialment.

En primer lloc, he après a fer recerca i a exemplificar, mitjançant un programa informàtic, les capacitats i les possibilitats que ofereix el PLN. Aleshores, he pogut constatar que encara hi ha molt de recorregut per investigar. Així doncs, aquest camp, conjuntament amb la revolució que experimentarà la Intel·ligència Artificial (IA) en un futur –segons els experts entrevistats-, evolucionarà i es convertirà definitivament en una eina d'ús quotidià imprescindible, tot i que actualment ja hi és present en molts aspectes.

En segon lloc, he pogut aprofundir i contrastar dos enfocaments complementaris de la Lingüística Computacional –un des d'un punt de vista informàtic i un altre, lingüístic-, gràcies a les entrevistes fetes als investigadors Lluís Padró i Montserrat Marimon, professors de la UPC i la UPF, respectivament. Després de conèixer-los i conversar amb ells fora d'entrevista i de manera més distesa, se m'ha despertat l'interès acadèmic que ja tenia latent de fa temps. També m'he adonat de la importància que té la informàtica en aquesta disciplina, però, alhora, de la necessitat de formar-se amb una bona base lingüística.

Al llarg de la recerca, he consolidat conceptes com ara els diferents nivells d'anàlisi de les oracions, els mètodes que existeixen per relacionar el llenguatge humà amb les màquines i les aplicacions que ofereix aquest camp d'estudi. He pogut comprovar que aquestes aplicacions es poden implementar en moltes disciplines, com ara la generació d'historials clínics en el cas de la medecina, la classificació de currículums en la secció de recursos humans de les empreses o la classificació d'incidències i antecedents penals al Departament d'Interior d'un govern. Concretament, en el meu cas i amb el programa informàtic amb què he treballat, he comprovat l'eficàcia del PLN en un àmbit actual i polèmic com són els mitjans de comunicació i la seva subjectivitat a l'hora de fer referència a temes polítics.

Finalment, l'elaboració d'aquest treball m'ha confirmat la proposta inicial sobre el meu futur acadèmic i professional. He decidit encarar els estudis universitaris enfocats en la LC i, per tant, en la Intel·ligència Artificial. He descobert un nou àmbit que m'apassiona

i, en conseqüència, m'iniciaré en estudis de programació per introduir-me parcialment a la recerca i la investigació del PLN.

6. Llista de referències

DATA JOBS. *Natural Language Processing* [en línia]. Accessible a https://datajobs.com/data-science-repo/NLP-Background-%5BSU%5D.pdf Consulta: 20/06/2018.

SCM. *Natural Language Processing (NLP)* [en línia]. Accessible a https://www.scm.tees.ac.uk/isg/aia/nlp/NLP-overview.pdf Consulta: 22/06/2018.

MT NEWS INTERNATIONAL. *Fifty years of the computer and translation* [en línia]. Accessible a http://www.hutchinsweb.me.uk/MTNI-16-1997.pdf Consulta: 12/07/2018.

SLD TRADOS. *What is Machine Translation?* [en línia]. Accessible a https://www.sdltrados.com/solutions/machine-translation/ Consulta: 12/07/2018.

A MEDIUM CORPORATION. *Natural Language Processing (NLP): Chomsky's Theories of Syntax* [en línia]. Accessible a https://medium.com/@ehfirst/natural-language-processing-nlp-chomskys-theories-of-syntax-92fb8fa3d035 Consulta: 13/07/2018.

FREELING 4.1 USER MANUAL. *FreeLing Tagset Description* [en línia]. Accessible a https://talp-upc.gitbook.io/freeling-4-1-user-manual/tagsets Consulta: 15/07/2018.

SEM1A5. *Morphological analysis* [en línia]. Accessible a https://www.cs.bham.ac.uk/~pjh/sem1a5/pt2/pt2_intro_morphology.html Consulta: 15/07/2018.

ARXIV. *Natural Language Processing: Sate of The Art, Current Trends and Challenges* [en línia]. Accessible a https://arxiv.org/ftp/arxiv/papers/1708/1708.05148.pdf Consulta: 15/07/2018.

EXPERT SYSTEM. *What is Natural Language Processing?* [en línia]. Accessible a https://www.expertsystem.com/ Consulta: 18/07/2018.

OMAR'S BRAIN. *Natural Language Processing – The Big Picture* [en línia]. Accessible a https://omarsbrain.wordpress.com/tag/natural-language-processing-linguistics-phonology-morphology-discourse-pragmatic-summarization/ Consulta: 20/07/2018.

REVISTA POLITÉCNICA. *Aplicaciones de Procesamiento de Lenguaje Natural* [en línia]. Accessible a https://rua.ua.es/dspace/bitstream/10045/33514/1/2013_Hernandez_Gomez_RevPolitec.pdf Consulta: 25/07/2018.

MICROSOFT. *Conceptos de mineria de datos* [en línia]. Accessible a https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017 Consulta: 25/07/2018.

SLIDESHARE. *Natural Language Processing* [en línia]. Accessible a https://www.slideshare.net/GirishKhanzode/nlp-52218202 Consulta: 27/07/2018.

SLIDESHARE. *Natural Language Processing Introduction* [en línia]. Accessible a https://www.slideshare.net/YasirAhmedKhan/natural-language-processing-42573796 Consulta: 27/07/2018.

EL PROFESIONAL DE LA INFORMACIÓN. *Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicacions (Parte I)* [en línia]. Accessible a http://www.elprofesionaldelainformacion.com/contenidos/1997/enero/procesamiento_del_lenguaje_natural_revisin_del_estado_actual_bases_tericas_y_aplicaci ones_parte_i.html Consulta: 15/08/2018.

INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT). *Natural language processin approaches, application and limitations* [en línia].

Accessible a file:///C:/Users/usuari/Downloads/IJERTV1IS7481.pdf Consulta: 16/08/2018.

FACULTAD DE PSICOLOGÍA – UNIVERSIDAD DE BUENOS AIRES. *Algunas ventajas del enfoque connexionista en cuanto al procesamiento del lenguaje natural* [en línia]. Accessible a https://www.aacademica.org/000-032/9.pdf Consulta: 17/08/2018.

EXPERT SYSTEM LAB. *Hybrid techniques for knowledge-based NLP* [en línia]. Accessible a http://expertsystemlab.com/kcap2017/ Consulta: 22/08/2018.

TECHNOPEDIA. *Natural Language Toolkit (NLTK)* [en línia]. Accessible a https://www.techopedia.com/definition/30343/natural-language-toolkit-nltk Consulta: 28/08/2018.

PYTHON SOFTWARE FOUNDATION. *Applications* [en línia]. Accessible a https://www.python.org/about/ Consulta: 03/09/2018.

FREELING HOME PAGE. *Welcome* [en línia]. Accessible a http://nlp.lsi.upc.edu/freeling/index.php/node/1 Consulta: 03/09/2018.

EXPERT SYSTEM LAB. *Tutorial on hybrid techinques for knowledge-based NLP* [en línia]. Accessible a http://expertsystemlab.com/hybridNLP18/ Consulta: 20/09/2018.

BSC. *Text mining* [en línia]. Accessible a https://www.bsc.es/ca/discover-bsc/organisation/scientific-structure/text-mining Consulta: 15/10/2018.

IIC. *Aplicaciones del procesamiento del lenguaje natural* [en línia]. Accessible a http://www.iic.uam.es/inteligencia/aplicaciones-procesamiento-lenguaje-natural/ Consulta: 15/10/2018

Annexos

ÍNDEX DE CONTINGUTS

1.1. Programa informàtic complet

```
import nltk, re, pprint, os, sys, math
from collections import defaultdict

# necesari per a la conexio al textserver
import requests
from xml.dom.minidom import parseString

## ---  gramatiques del chunker ----
gramatica = {}

## -- definicio regles espanyol
gramatica["es"] = r"""
        N: {<NC.*|NP.*>}
        SA: {<RG>*<AQ.*|VMP.*>}
        SV: {<VM.*><D.*><N>}
        SN:{<N><SP.*><D.*>?<N>}
            {<D.*>?<SA>+<N>+}
            {<D.*>?<N>+<SA>+<N>?}
        """

## -- definicio regles catala
gramatica["ca"] = r"""
        N: {<NC.*|NP.*>}
        SA: {<RG>*<AQ.*|VMP.*>}
        SV: {<VM.*><D.*><N>}
        SN:{<N><SP.*><D.*>?<N>}
            {<D.*>?<SA>+<N>+}
            {<D.*>?<N>+<SA>+<N>?}
            """
```

Il·lustració 1: *Fragment 1 del programa informàtic analitzador de notícies.*

FONT: Elaboració pròpia

```
## -- definicio regles anglès
gramatica["en"] = r"""
       SV: {<VB.?><DT><NNPS?|NNPS?>}
       SN: {<NNS?|NNPS?><IN><DT>?<NNS?|NNPS?>}
           {<JJ.?>+<NNS?|NNPS?>+}
           {<NNS?|NNPS?><NNS?|NNPS?>+}
       SA: {<RB.?>*<JJ.?>}
        N: {<NNS?|NNPS?>}
        """

# funcio que envia el text a analitzar al textserver
def freeling(lang, text) :

    # Create request
    request_data = {'username':'jofre',
                    'password':'jofre.123',
                    'text_input':text,
                    'language':lang,
                    'output':'xml',
                    'interactive':'1' }
    url = "http://frodo.lsi.upc.edu:8080/TextWS/textservlet/ws/processQuery/tagger"
    # Send request and get response
    resp = requests.post(url, files=request_data)

    # process response (posar-la en el format adqueat pel RegexpParser)
    xml = parseString(resp.text)
    result = []
    for s in xml.getElementsByTagName("sentence"):
        tks = []
        for t in s.getElementsByTagName("token") :
            tks.append((t.attributes["form"].value, t.attributes["tag"].value))

        result.append(tks)

    return result
```

Il·lustració 2: *Fragment 2 del programa informàtic analitzador de notícies.*

FONT: Elaboració pròpia

```
# Funció que analitza els textos en l'idioma demanat.
# Busca els textos al directori noticies-<lang>  (noticies-ca, noticies-es)
# deixa els resultats al directori noticies-<lang>-out  (noticies-ca-out, noticies-es-out)
def analitza(lang, dir_in, minaparicions, result):

    aparicionstotal = defaultdict(int)  # num aparicions totals de cada paraula
    aparicionsdiari = {}                # num aparicions de cada paraula en cada diari
    longituddiari = defaultdict(int)    # num de paraules en les noticies d'un diari
    longitudtotal = 0.0                 # num de paraules en totels les noticies

    dir_out = dir_in+"-out"   # directori on deixar les paraules extretes

    for f in os.listdir(dir_in) :
        print ("tractant fitxer"+dir_in+"/"+f)
        # especifica els fitxers d'entrada i de sortida
        noticia = open(dir_in+"/"+f, encoding='utf-8-sig').read()
        sortida = open(dir_out+"/"+f, 'w')

        # nom del diari
        nomdiari = f.split("_")[0]
        if nomdiari not in aparicionsdiari : aparicionsdiari[nomdiari] = defaultdict(int)

        # fer el tagging de la noticia
        frases_etiquetades = freeling(lang, noticia)

        # gramàtica i parser (analitzador sintàctic)
        parser = nltk.RegexpParser(gramatica[lang])

        # per a cadascuna de les frases etiquetades  aplicar les regles del parser
        for sent in frases_etiquetades:
            tree = parser.parse(sent)

            for subtree in tree.subtrees():
                # copia els subarbres reconeguts per la gramatica al fitxer de sortida,
                # i incrementa els comptadors.
                if (subtree.label() != 'S') :
                    print(subtree,  file=sortida)
```

Il·lustració 3: *Fragment 3 del programa informàtic analitzador de notícies.*

FONT: Elaboració pròpia

```
111                     print(subtree,  file=sortida)
112                     s = ""
113                     for x in subtree.leaves() : s += x[0].lower()+"_"
114                     s = s[:-1]
115                     aparicionsdiari[nomdiari][s] += 1
116                     aparicionstotal[s] += 1
117                     longituddiari[nomdiari] += 1
118                     longitudtotal += 1.0
119
120         sortida.close()
121
122     #  Calcular TF-IDF  ----------------------------------
123     idf = {}
124     numdiaris = {}
125     for w in aparicionstotal :
126         numdiaris[w] =0
127         for d in aparicionsdiari :
128             if w in aparicionsdiari[d] :
129                 numdiaris[w] += 1
130         idf[w] = math.log(1.0*len(aparicionsdiari)/numdiaris[w])/math.log(2)
131
132     # escriure resultats TF-IDF
133     fitxerTFIDF = open(result+"-TFIDF.txt", 'w')
134     for d in sorted(aparicionsdiari) :
135         print("--------------------",d,"----------------------", file=fitxerTFIDF)
136         print ("ApD ApT  ND   IDF TFIDF paraula", file=fitxerTFIDF)
137         for w in sorted(aparicionsdiari[d], key=lambda w : aparicionsdiari[d][w]*idf[w], reverse=True) :
138             if aparicionstotal[w]>minaparicions :
139                 print ("%3d %3d %3d %5.2f %5.2f %s" % (aparicionsdiari[d][w],
140                 aparicionstotal[w],numdiaris[w],idf[w],aparicionsdiari[d][w]*idf[w],w), file=fitxerTFIDF)
141                 fitxerTFIDF.close()
```

Il·lustració 4: *Fragment 4 del programa informàtic analitzador de notícies.*

FONT: Elaboració pròpia

```
144     # Calcular Informacio Mutua ------------------------------------------------
145     im = {}
146     for w in aparicionstotal :
147         for d in aparicionsdiari :
148             if d not in im : im[d] = defaultdict(float)
149             if w in aparicionsdiari[d] :
150                 im[d][w] = math.log((aparicionsdiari[d][w]*longitudtotal)/
151                   (aparicionstotal[w]*longituddiari[d]))/math.log(2)
152
153     # escriure resultats IM
154     fitxerIM = open(result+"-IM.txt", 'w')
155     for d in sorted(aparicionsdiari) :
156         print("--------------------",d,"----------------------", file=fitxerIM)
157         print ("ApD ApT  NPD  NPT    IM paraula", file=fitxerIM)
158         for w in sorted(aparicionsdiari[d], key=lambda w : im[d][w], reverse=True) :
159             if aparicionstotal[w]>minaparicions :
160                 print ("%3d %3d %4d %4.0f %5.2f %s" % (aparicionsdiari[d][w],aparicionstotal[w],
161                                                        longituddiari[d],longitudtotal,im[d][w],w),
162                                                         file=fitxerIM)
163     fitxerIM.close()
```

Il·lustració 5: *Fragment 5 del programa informàtic analitzador de notícies.*

FONT: Elaboració pròpia

1.2. Entrevistes a experts

1.2.1. Entrevista a Lluís Padró

En quin punt de la teva vida professional decideixes entrar al món de la lingüística computacional?

"Quan vaig acabar la carrera vaig treballar un parell d'anys en una empresa de software, però, al cap del temps, notava la necessitat d'aprendre coses noves i vaig decidir fer un doctorat. Entre els temes que m'interessaven hi havia la Intel·ligència Artificial (IA), particularment el Processament del Llenguatge Natural (PLN), ja que sempre m'han encuriosit les estructures de la llengua."

Parteixes d'una enginyeria informàtica com a base de la teva formació professional. De quina manera accedeixes al camp de la lingüística?

"Jo no diria que treballo en el camp de la lingüística. Jo faig PLN, que és una branca de la IA que intenta que les màquines processin el llenguatge humà. Per fer-ho, és evident que fan falta coneixements lingüístics, però sovint s'usen mètodes estadístics en els quals la part lingüística es troba en la codificació de les dades, cosa que solen fer lingüistes.

Una cosa és l'enginyeria (que permet fer màquines útils) i l'altra la ciència (que permet explicar com és el món). Fixa't que això no és una contradicció i que passa sovint a la història: La catapulta es va inventar abans de conèixer la llei de la gravetat i la fórmula del tir parabòlic. La combinació química de compostos (per crear sabó, per exemple) es feia molt abans de conèixer la teoria atòmica sobre com es combinen les molècules.

La diferència rau en què l'objectiu de la lingüística és explicar com funciona el llenguatge (i el de la lingüística computacional és explicar-ho d'una forma representable en un ordinador), mentre que l'objectiu del PLN és l'execució de màquines que processin la llengua, independentment de si el model que usen explica com aquesta funciona o no. És a dir, si la LC[16] genera un model computable de la llengua, el PLN el podrà usar i el problema estarà resolt. Mentrestant, fem servir tot el que podem (estadística, xarxes neuronals, etc.) que produeixi resultats útils, encara que sabem que el llenguatge no funciona realment així."

[16] Lingüística Computacional

De quina forma preveus que evolucioni aquesta disciplina des d'un punt de vista informàtic?

"La tendència que es preveu actualment és que les xarxes neuronals portin, en un futur proper, una revolució a la IA que resoldrà molts d'aquests problemes o, almenys, des d'un punt de vista pràctic. És a dir, les màquines podran aprendre el llenguatge de forma similar a la que ho fa un nen i, d'aquesta manera, ja no seran necessaris models lingüístics per fer que les màquines ens entenguin.

Jo personalment crec que no serà tan imminent com diuen i que haurem de seguir amb els mètodes actuals força temps. D'altra banda, si fos el cas, tindríem una màquina útil, però encara no hauríem respòs la pregunta de «com funciona el llenguatge?», dit d'una altra manera: tindríem una catapulta, però encara no sabríem la fórmula del tir parabòlic."

Aquesta professió trenca l'esquema tradicional de la separació entre carreres de lletres i de ciències?

"Sí i no... El trenca perquè tradicionalment hem posat la lingüística a «lletres», ja que està relacionada amb la llengua. A més a més, tradicionalment, el mètode usat pels lingüistes «clàssics» ha estat més de lletres que de ciències.

Tanmateix, jo crec que hauríem de veure la lingüística (o almenys la computacional) com una disciplina de «ciències»: la Lingüística Computacional crea models formals (matemàtics) de la llengua i en contrasta la validesa experimentalment. És a dir, segueix el mateix mètode científic que qualsevol altra ciència.

Per tant, en comptes de trencar l'esquema, jo crec que el que fa és posar la lingüística al lloc que li correspon."

La lingüística computacional està formada per dues disciplines: la lingüística i la informàtica. Quin d'aquests dos àmbits creus que hi juga un paper més fonamental?

"Com deia abans, depèn de l'objectiu en què posis l'èmfasi: si vols màquines útils que processin la llengua, té mes protagonisme la informàtica. Si, en canvi, vols models que expliquin com funciona la llengua i perquè és com és, llavors és més important la lingüística."

Dins dels estudis d'enginyeria informàtica, creus que és necessari que s'inclogui una base de formació lingüística?

“L’enginyeria informàtica és molt àmplia i no a tothom li interessaria aquest tema.

Jo ho posaria dins d’una assignatura de PLN, de manera que qui estigui interessat en el tema tingui els coneixements bàsics.”

1.2.2. Entrevista a Montserrat Marimon

En quin punt de la seva vida professional decideix entrar al món de la lingüística computacional?

Jo vaig estudiar filologia anglesa. Quan estava acabant la carrera, va sortir un màster en lingüística computacional. Em va interessar i m’hi vaig apuntar.

Parteix d’una carrera de lletres pures. De quina manera accedeix al camp de la programació i la informàtica?

A la primera feina que vaig tenir en un projecte de recerca a la UB, em van donar una màquina Unix (de les d’abans, sense finestres ni entorn gràfic) i un sistema de gramàtiques, i em van dir «espavila’t per fer una gramàtica del castellà». Em vaig anar espavilant i aprenent Unix i a fer gramàtiques d’aquestes (que en el fons es similar a un llenguatge de programació). Més tard, he après a programar en Python fent cursos online.

Amb l’experiència d’ara, tornaria a fer el mateix recorregut acadèmic o prioritzaria els estudis informàtics a l’inici?

Ara faria lingüística i no filologia anglesa (abans no existia la titulació de lingüística). També faria els cursos online o aprendria a programar abans. El que no faria és estudiar el grau d’informàtica.

Quina projecció laboral abraça la lingüística computacional?

Fins fa poc temps en tenia poca. Bàsicament treballar en projectes de recerca a la universitat o en alguna empresa dedicada a desenvolupar diccionaris electrònics o algun dels primers traductors automàtics.

Però d’uns anys ençà hi ha moltes més oportunitats, Hi ha força empreses que desenvolupen eines relacionades amb la llengua: traductors automàtics, *chatbots* d’atenció al client, gestors de documents... Sovint aquestes empreses es basen només en la part informàtica, però mica en mica van veient que incorporar lingüistes als seu desenvolupament resulta en millors productes.

Aquesta disciplina està dividida en subcamps?

La lingüística computacional ja és un subcamp de la lingüística. Subdividir-ho encara més resultaria en «microcamps» sense una perspectiva global. Si que hi ha diferents àrees d'aplicació (traducció automàtica, anàlisi sintàctica i/o semàntica, etc.), però no arriben a ser subcamps perquè estan molt relacionats entre si.

De quina forma preveu que evolucioni la lingüística computacional?

Sembla que les xarxes neuronals basades en *Deep Learning* han de menjar-s'ho tot. Si això acaba sent així, el paper dels lingüistes quedarà més restringit a estructurar i codificar les dades necessàries per entrenar les xarxes neuronals. També quedarà l'àmbit acadèmic, investigant models sobre com funciona la llengua, però sense una necessitat real d'aplicar aquests models teòrics a la pràctica. Si les xarxes neuronals al final no són tan la "panacea" com diuen (cosa que no seria pas el primer cop que passa), llavors probablement l'escenari serà millor, perquè caldrà tornar als models teòrics per resoldre les tasques que les xarxes no hagin pogut.

Creu que es necessària la creació d'uns estudis dedicats específicament a la lingüística computacional?

Si, és clar. Seria molt interessant!

Aquesta professió trenca l'esquema tradicional de la separació entre carreres de lletres i de ciències?

És cert que la lingüística sempre ha estat encasellada en àrees de lletres, i que la gent que l'estudia sol tenir un *background* d'humanitats. Però la lingüística no és realment una disciplina de lletres. Hi ha molta base formal i models teòrics matemàtics. La part experimental, que és el que li faltava per poder ser una ciència «dura», és cada cop més sòlida gràcies a les grans quantitats de dades de la llengua (textos i àudios en múltiples idiomes) que estan esdevenint disponibles gràcies a Internet.

D'altra banda, en general, diferents subàrees de la lingüística toquen amb diferents camps del coneixement: la neurolingüística es relaciona amb la neurologia per estudiar com el cervell processa la llengua, la sociolingüística amb la sociologia, l'antropologia lingüística, que estudia la relació entre la llengua i la cultura, i la lingüística computacional amb la informàtica.

1.3. Taules

1.3.1. Diaris originalment en castellà

1.3.1.1. ABC

ABC							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	**3**	**3.62**	**censuren_l'_ofensiva**	1	31	-1.34	lamela
3	3	3.62	govern_de_sánchez	**1**	**32**	**-1.38**	**la_justícia_alemanya**
3	**3**	**3.62**	**desdeny**	**1**	**32**	**-1.38**	**sedició**
3	**3**	**3.62**	**l'_expresident_cessat**	**1**	**33**	**-1.43**	**fuga**
3	**3**	**3.62**	**una_resposta_ contundent**	1	34	-1.47	cap
3	3	3.62	fórmules	1	35	-1.51	madrid
3	3	3.62	espadaler	**1**	**37**	**-1.59**	**eleccions**
6	**7**	**3.39**	**secessionisme**	**1**	**39**	**-1.67**	**membres**
5	6	3.35	proposta	**1**	**39**	**-1.67**	**parlament**
4	**5**	**3.29**	**comissió_permanent**	**5**	**198**	**-1.69**	**president**
4	5	3.29	partida	1	41	-1.74	part
4	5	3.29	símbols	**1**	**48**	**-1.97**	**referèndum**
3	**4**	**3.20**	**vigilada**				
3	**4**	**3.20**	**contundent**				
3	**4**	**3.20**	**denunciada**				
4	6	3.03	abc				
2	**3**	**3.03**	**rebel**				
2	**3**	**3.03**	**insta_el_govern**				
2	3	3.03	aire				
2	**3**	**3.03**	**secessionistes**				
2	**3**	**3.03**	**defensa_de_el_magistrat**				
2	3	3.03	encarregat_de _nomenar-seria				
2	3	3.03	nomenar-seria				
2	**3**	**3.03**	**ambaixador_d'_espanya**				
2	**3**	**3.03**	**país_en_qüestió**				
2	3	3.03	exemple				
2	**3**	**3.03**	**fundacional**				

Taula 1: *Taula elaborada amb els resultats de les notícies de l'*ABC *amb els valors de la mesura IM.*

1.3.1.2. EL MUNDO

EL MUNDO							
VALORS POSITIUS				VALORS NEGATIUS			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	3	3.64	continuïtat	**1**	**24**	**-0.94**	**president_de_la_generalitat**
3	3	3.64	immensa	1	26	-1.06	sánchez
3	**3**	**3.64**	**feblesa**	1	26	-1.06	quim_torra
4	5	3.32	demostració	**1**	**26**	**-1.06**	**catalana**
2	**3**	**3.06**	**detenció_europea**	1	27	-1.11	premsa
2	**3**	**3.06**	**reclamats**	**1**	**27**	**-1.11**	**traïció**
2	**3**	**3.06**	**el_líder_sobiranista**	**1**	**27**	**-1.11**	**suport**
2	3	3.06	exercit	**1**	**27**	**-1.11**	**el_govern_espanyol**
2	3	3.06	començament	1	29	-1.22	alta
2	**3**	**3.06**	**tribunals_de_bèlgica**	1	30	-1.27	temps
2	**3**	**3.06**	**conseqüència**	1	30	-1.27	situació
2	**3**	**3.06**	**instrument_polític**	**1**	**31**	**-1.31**	**waterloo**
6	10	2.90	indicat	**1**	**32**	**-1.36**	**pablo_llarena**
3	**5**	**2.90**	**militants**	**1**	**33**	**-1.40**	**fuga**
6	**11**	**2.77**	**sobiranista**	**1**	**35**	**-1.49**	**madrid**
2	**4**	**2.64**	**la_fiscalia_espanyola**	**1**	**35**	**-1.49**	**acord**
2	**4**	**2.64**	**l'_ex_president**	**1**	**38**	**-1.61**	**risc**
2	4	2.64	freda	**1**	**44**	**-1.82**	**llei**
2	4	2.64	germànic	**1**	**53**	**-2.09**	**consellers**
2	**4**	**2.64**	**rumb_a_bèlgica**	1	65	-2.38	fet
2	**4**	**2.64**	**divisió**				
2	**4**	**2.64**	**moral**				
3	6	2.64	instrument				
2	**4**	**2.64**	**hostil**				
4	**9**	**2.47**	**asil**				
3	**7**	**2.42**	**rumb**				
3	**7**	**2.42**	**crida_nacional_per_la_república**				
3	**7**	**2.42**	**convergència**				
2	**5**	**2.32**	**legals**				
2	5	2.32	consultats				
2	**5**	**2.32**	**víctima**				
2	5	2.32	figura				
3	8	2.23	experts				
1	3	2.06	rtbf				
1	**3**	**2.06**	**desobediència_a_l'_autoritat**				

Taula 2: *Taula elaborada amb els resultats de les notícies de* El Mundo *amb els valors de la mesura IM.*

1.3.1.3. EL PAÍS

EL PAÍS							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
4	4	2.86	arbre	**1**	**19**	**-1.38**	**ordre_de_detenció**
3	**3**	**2.86**	**eurodiputats**	**1**	**19**	**-1.38**	**diputat**
4	**4**	**2.86**	**els_nacionalistes_flamencs**	1	19	-1.38	països
4	4	2.86	flamencs	**1**	**19**	**-1.38**	**custòdia**
3	**3**	**2.86**	**autònoma**	**3**	**59**	**-1.43**	**defensa**
4	4	2.86	heger	**1**	**20**	**-1.46**	**la_policia_alemanya**
5	**5**	**2.86**	**interns**	1	20	-1.46	any
3	3	2.86	dissolució	1	20	-1.46	cop
5	5	2.86	editorial	1	21	-1.53	lluís_puig
3	**3**	**2.86**	**populista**	1	21	-1.53	forma
4	**5**	**2.54**	**parlament_europeu**	**1**	**21**	**-1.53**	**resolució**
4	**5**	**2.54**	**legislatura**	**2**	**43**	**-1.56**	**mesures**
3	4	2.45	preguntes	1	22	-1.60	mitjans
3	**4**	**2.45**	**crítiques**	**1**	**22**	**-1.60**	**resposta**
3	4	2.45	previst	**1**	**23**	**-1.66**	**independent**
3	**4**	**2.45**	**la_llei_alemanya**	**1**	**23**	**-1.66**	**separatista**
5	7	2.38	important	1	24	-1.72	termini
2	**3**	**2.28**	**el_govern_destituït**	**1**	**24**	**-1.72**	**frontera**
2	**3**	**2.28**	**sobiranisme**	1	25	-1.78	twitter
2	3	2.28	sostingut	**1**	**26**	**-1.84**	**catalana**
2	**3**	**2.28**	**comicis**	**1**	**27**	**-1.89**	**executiu**
2	**3**	**2.28**	**aplicar_l'_article**	1	28	-1.94	finlàndia
2	3	2.28	bateria	**1**	**33**	**-2.18**	**drets**
2	**3**	**2.28**	**pau**	**1**	**35**	**-2.27**	**acord**
2	3	2.28	magnitud	**1**	**47**	**-2.69**	**república**
2	3	2.28	autor	**1**	**128**	**-4.14**	**llarena**
2	3	2.28	prorrogables				
2	**3**	**2.28**	**vaga**				
2	3	2.28	jove				
2	3	2.28	documents				
2	3	2.28	despertat				
2	3	2.28	curiositat				
2	**3**	**2.28**	**el_polític_català**				
4	**6**	**2.28**	**cel·la**				
2	3	2.28	visita				
2	**3**	**2.28**	**espera_de_judici**				
2	**3**	**2.28**	**jugada**				

Taula 3: *Taula elaborada amb els resultats de les notícies de* El País *amb els valors de la mesura IM.*

1.3.1.4. LA RAZÓN

LA RAZÓN							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	**3**	**3.81**	**atemptat**	**2**	**59**	**-1.07**	**polítics**
3	**3**	**3.81**	**justicia**	1	31	-1.14	ciutadans
3	3	3.81	felip	1	31	-1.14	waterloo
3	3	3.81	protagonisme	**1**	**32**	**-1.19**	**la_justícia_alemanya**
3	**3**	**3.81**	**implantar_la_república**	1	32	-1.19	comunicat
3	**4**	**3.39**	**actuació_de_l'_advocacia _de_l'_estat**	**1**	**32**	**-1.19**	**pablo_llarena**
3	4	3.39	ramon_espadaler	**2**	**66**	**-1.23**	**euroordre**
3	**4**	**3.39**	**víctimes**	**1**	**33**	**-1.23**	**drets**
5	**7**	**3.32**	**rei**	1	35	-1.32	regional
2	**3**	**3.22**	**ordre_d'_arrest**	**1**	**35**	**-1.32**	**madrid**
2	3	3.22	neerlandès	**3**	**110**	**-1.39**	**catalunya**
2	3	3.22	convertit	**3**	**110**	**-1.39**	**fiscalia**
2	**3**	**3.22**	**decisió_de_el_tribunal**	**1**	**37**	**-1.40**	**dret**
2	**3**	**3.22**	**portada_a_terme**	**1**	**42**	**-1.58**	**internacional**
2	3	3.22	el_secretari_general	**1**	**46**	**-1.71**	**pdecat**
2	3	3.22	cdr	**1**	**48**	**-1.78**	**referèndum**
2	3	3.22	famílies	**5**	**255**	**-1.86**	**alemanya**
2	3	3.22	pols	**1**	**52**	**-1.89**	**detingut**
2	3	3.22	tarragona	**1**	**53**	**-1.92**	**consellers**
2	3	3.22	fitxatge	**1**	**57**	**-2.02**	**delictes**
2	3	3.22	alcalde				
3	**5**	**3.07**	**tasca_jurisdiccional**				
3	5	3.07	pilar_rahola				
4	**7**	**3.00**	**veredicte**				

Taula 4: *Taula elaborada amb els resultats de les notícies de* La Razón *amb els valors de la mesura IM.*

1.3.1.5. LA VANGUARDIA

LA VANGUARDIA							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	3	3.51	braç	**1**	**23**	**-1.01**	**preventiva**
3	**3**	**3.51**	**una_catalunya_independent**	1	23	-1.01	dit
3	**4**	**3.10**	**el_president_cessat**	**1**	**23**	**-1.01**	**candidat**
3	4	3.10	donada	1	24	-1.07	acte
3	**4**	**3.10**	**lliurament_a_espanya**	**1**	**24**	**-1.07**	**barcelona**
3	**4**	**3.10**	**el_mateix_puigdemont**	1	24	-1.07	majoria
3	**4**	**3.10**	**sala_segona_del_tribunal_suprem**	1	24	-1.07	meritxell_serret
2	**3**	**2.93**	**resta_de_exconsellers**	**2**	**48**	**-1.07**	**líder**
2	3	2.93	josé_manuel_maza	1	24	-1.07	empara
2	3	2.93	mesa_del_parlament	**8**	**198**	**-1.12**	**president**
2	**3**	**2.93**	**exposat**	1	25	-1.13	instructor
2	**3**	**2.93**	**resta_de_consellers**	**2**	**51**	**-1.16**	**catalans**
2	**3**	**2.93**	**gravetat**	**3**	**77**	**-1.17**	**brussel·les**
2	3	2.93	actuat	1	26	-1.19	instrucció
2	3	2.93	neumuenster				
2	**3**	**2.93**	**competent**				
2	3	2.93	cabals				
2	**3**	**2.93**	**fiscalia_general_de_schleswig-holstein**				
2	**3**	**2.93**	**desplaçats**				
2	**3**	**2.93**	**el_govern_alemany**				
2	3	2.93	pasqua				
2	3	2.93	lleida				
4	**6**	**2.93**	**reglament**				
2	**3**	**2.93**	**llibertat_d'_expressió**				

Taula 5: *Taula elaborada amb els resultats de les notícies de* La Vanguardia *amb els valors de la mesura IM.*

1.3.1.6. LIBERTAD DIGITAL

LIBERTAD DIGITAL							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	3	3.58	versió	**1**	**33**	**-1.47**	**drets**
3	3	3.58	mostres	1	34	-1.51	cap
4	5	3.26	amo	**1**	**35**	**-1.55**	**madrid**
3	**4**	**3.16**	**posada_en_llibertat**	1	38	-1.67	nou
3	**4**	**3.16**	**fugitiu**	1	39	-1.71	membres
3	**4**	**3.16**	**excel·lent**	3	117	-1.71	decisió
5	7	3.09	cuevillas	**1**	**41**	**-1.78**	**europea**
2	3	2.99	enfocats	**1**	**42**	**-1.81**	**internacional**
2	**3**	**2.99**	**cop_d'_estat**	1	42	-1.81	mateix
2	3	2.99	conegut	**1**	**42**	**-1.81**	**la_justícia_espanyola**
4	**6**	**2.99**	**autonòmic**	**2**	**89**	**-1.90**	**país**
2	3	2.99	diaris	1	47	-1.98	hores
2	**3**	**2.99**	**ressò**	**1**	**47**	**-1.98**	**independentista**
2	3	2.99	bones	**1**	**48**	**-2.01**	**tribunal_suprem**
2	**3**	**2.99**	**escapada**	**1**	**48**	**-2.01**	**líder**
4	6	2.99	alonso_cuevillas	**2**	**96**	**-2.01**	**procés**
4	6	2.99	mansió	**1**	**57**	**-2.26**	**delictes**
2	3	2.99	elsa_artadi	**2**	**139**	**-2.54**	**tribunal**
2	**3**	**2.99**	**iniciativa**	1	77	-2.69	cas
4	**6**	**2.99**	**assetjament**				
2	3	2.99	identificat				
2	**3**	**2.99**	**urgent**				
2	**3**	**2.99**	**gravíssim**				
2	3	2.99	simbologia				
2	3	2.99	comercial				
7	11	2.93	llaços				
11	**18**	**2.87**	**separatistes**				
3	5	2.84	posada				
7	**13**	**2.68**	**fugat**				
12	**23**	**2.64**	**separatista**				
2	**4**	**2.58**	**intervingut**				
2	**4**	**2.58**	**pena_de_presó**				

Taula 6: *Taula elaborada amb els resultats de les notícies de* Libertad Digital *amb els valors de la mesura IM.*

1.3.1.7. OK DIARIO

OK DIARIO							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	3	3.74	dolors_bassa	**1**	**27**	**-1.01**	**querella**
3	**3**	**3.74**	**exconseller**	**5**	**147**	**-1.13**	**espanyol**
3	**3**	**3.74**	**fugida_de_el_colpista**	1	31	-1.21	general
8	**8**	**3.74**	**colpista**	**1**	**32**	**-1.26**	**la_justícia_alemanya**
3	**3**	**3.74**	**falsificat**	**1**	**32**	**-1.26**	**civil**
3	3	3.74	oral	**4**	**135**	**-1.33**	**extradició**
4	**5**	**3.42**	**seguiment**	**4**	**139**	**-1.38**	**tribunal**
4	**5**	**3.42**	**balisa**	1	35	-1.39	setmana
2	**3**	**3.16**	**exconsellers_fugats**	1	35	-1.39	portaveu
2	**3**	**3.16**	**desobediència_a_ l'_autoritat**	1	36	-1.43	càrrec
4	6	3.16	oede	1	36	-1.43	passat
2	3	3.16	meritxell_borràs	**1**	**37**	**-1.47**	**eleccions**
2	3	3.16	treball	1	38	-1.50	nou
2	3	3.16	santi_vila	1	39	-1.54	membres
2	**3**	**3.16**	**reiterades**	**5**	**198**	**-1.56**	**president**
2	**3**	**3.16**	**previsible**	**1**	**40**	**-1.58**	**independentisme**
2	3	3.16	viatjat	**1**	**42**	**-1.65**	**presos**
2	3	3.16	feina	**1**	**43**	**-1.68**	**lliurament**
2	**3**	**3.16**	**impunitat**	1	45	-1.75	erc
2	3	3.16	habituals	**1**	**51**	**-1.93**	**catalans**
2	**3**	**3.16**	**el_sistema_europeu**	**1**	**55**	**-2.04**	**partit**
2	**3**	**3.16**	**cooperació_policial**	**1**	**72**	**-2.43**	**violència**
2	3	3.16	mida				
2	3	3.16	camp				
2	3	3.16	currículum				
2	3	3.16	organismes				
2	**3**	**3.16**	**conselleria_d'_interior**				
2	**3**	**3.16**	**serveis_de_seguretat**				
7	11	3.09	document				
5	8	3.07	europa_press				
3	5	3.01	formulada				
3	**5**	**3.01**	**incondicional**				
6	**11**	**2.87**	**colpistes**				

Taula 7: *Taula elaborada amb els resultats de les notícies de l'*OK Diario *amb els valors de la mesura IM.*

1.3.1.8. EL NACIONAL

EL NACIONAL							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	3	4.62	el_nacional	**1**	**33**	**-0.42**	**fuga**
3	**3**	**4.62**	**president_a_l'_exili**	**2**	**66**	**-0.42**	**euroordre**
4	**5**	**4.30**	**desplaçament**	1	34	-0.46	cap
3	4	4.21	teresa_cunillera	**1**	**35**	**-0.51**	**estranger**
7	**10**	**4.11**	**expresidents**	**3**	**110**	**-0.57**	**fiscalia**
2	**3**	**4.04**	**delegada_de_el_govern**	**1**	**38**	**-0.62**	**risc**
2	3	4.04	jordi_borràs	1	38	-0.62	anys
2	3	4.04	empreses	1	38	-0.62	nou
2	**3**	**4.04**	**desplaçament_d'_agents**	**1**	**40**	**-0.70**	**independentisme**
2	**3**	**4.04**	**policia_de_catalunya**	**2**	**82**	**-0.73**	**alemany**
2	3	4.04	altres_consideracions	**1**	**41**	**-0.73**	**europea**
2	**3**	**4.04**	**retribucions**	**1**	**42**	**-0.77**	**la_justícia_espanyola**
2	3	4.04	impedit	**1**	**42**	**-0.77**	**advocats**
2	**3**	**4.04**	**exercici_de_els_drets**	**1**	**43**	**-0.80**	**lliurament**
2	**3**	**4.04**	**cívics**	**2**	**89**	**-0.85**	**país**
2	3	4.04	ministres	**1**	**46**	**-0.90**	**independentistes**
2	**3**	**4.04**	**impugnada**	**1**	**46**	**-0.90**	**causa**
2	**3**	**4.04**	**cap_partit_polític**	**2**	**96**	**-0.96**	**procés**
2	**3**	**4.04**	**govern_davant_la_junta _electoral_central**	**1**	**48**	**-0.96**	**líder**
2	**3**	**4.04**	**els_drets_fonamentals**	1	52	-1.08	públics
2	**3**	**4.04**	**vetar_la_investidura**	1	55	-1.16	fons
8		3.92	escorta	**2**	**114**	**-1.21**	**català**
3	5	3.89	tuit	**1**	**57**	**-1.21**	**delictes**
3	5	3.89	reconeguts	**3**	**174**	**-1.23**	**expresident**
4	**7**	**3.82**	**impedir_la_investidura**	**3**	**176**	**-1.25**	**espanya**
2	**4**	**3.62**	**fonts_de_la_fiscalia**	1	61	-1.31	neumünster
2	**4**	**3.62**	**el_jutge_alemany**	**3**	**192**	**-1.38**	**presó**
2	4	3.62	guàrdia	1	65	-1.40	fet
2	4	3.62	escriptor	**2**	**140**	**-1.51**	**estat**
2	4	3.62	argumentació	**1**	**72**	**-1.55**	**exconsellers**
2	**4**	**3.62**	**ministeri_d'_afers_exteriors**	**1**	**77**	**-1.64**	**brussel·les**
2	4	3.62	receptor	1	91	-1.88	dies
2	4	3.62	consideracions	**1**	**117**	**-2.25**	**detenció**
2	**4**	**3.62**	**els_serveis_jurídics**	1	117	-2.25	decisió
3	6	3.62	departament				
2	4	3.62	competències				
2	4	3.62	activitat				
3	6	3.62	distància				
3	6	3.62	electorals				
2	**4**	**3.62**	**junta_electoral_central**				
3	**6**	**3.62**	**querellats**				
2	**4**	**3.62**	**juristes**				

Taula 8. *Taula elaborada amb els resultats de les notícies de* El Nacional *amb els valors de la mesura IM.*

1.3.2. Diaris originalment en català

1.3.2.1. ARA

ARA							
VALOR POSITIU				**VALOR NEGATIU**			
APD	**APT**	**IM**	**PARAULA**	**APD**	**APT**	**IM**	**PARAULA**
3	**3**	**3.26**	**consellers_a_brussel·les**	**1**	**23**	**1.26**	**preventiva**
3	**3**	**3.26**	**ordre_de_recerca**	**1**	**23**	**1.26**	**independent**
4	**4**	**3.26**	**captura_internacional**	1	24	1.32	acte
4	4	3.26	interlocutòries	1	25	1.38	grup
4	4	3.26	els_pròxims_dies	**1**	**25**	**1.38**	**advocacia_de_l'_estat**
3	**3**	**3.26**	**il·lícita**	**1**	**25**	**1.38**	**l'_estat_espanyol**
4	4	3.26	aportades	1	26	1.44	quim_torra
3	3	3.26	döpper	**1**	**27**	**1.49**	**premsa**
3	**3**	**3.26**	**causa_contra_el_procés**	**1**	**27**	**1.49**	**querella**
4	**4**	**3.26**	**poder_judicial**	4	114	1.57	català
3	**3**	**3.26**	**exconsellers_a_l'_exili**	1	29	1.59	possible
4	5	2.94	pròxims	**1**	**31**	**1.69**	**tribunals**
3	**4**	**2.85**	**part_de_el_govern**	**1**	**31**	**1.69**	**ciutadans**
6	**8**	**2.85**	**audiència_territorial**	**1**	**32**	**1.74**	**la_justícia_alemanya**
6	8	2.85	ara				
3	**4**	**2.85**	**proves**				
3	**4**	**2.85**	**transversal**				
5	**7**	**2.78**	**costos**				
5	7	2.78	pròximes				
2	**3**	**2.68**	**el_ministeri_públic**				
2	3	2.68	interpol				
2	**3**	**2.68**	**comú_acord**				
2	**3**	**2.68**	**alçament**				
2	3	2.68	organitzat				
2	3	2.68	associacions				
2	**3**	**2.68**	**la_policia_federal_belga**				
4	6	2.68	anunci				
2	**3**	**2.68**	**cas_de_puigdemont**				
2	**3**	**2.68**	**el_tribunal_regional**				
2	**3**	**2.68**	**demanda_de_puigdemont**				

Taula 9: *Taula elaborada amb els resultats de les notícies de l'*Ara *amb els valors de la mesura IM.*

1.3.2.2. EL PERIÓDICO

EL PERIÓDICO							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	3	4.14	directius	**3**	**114**	**-1.11**	**català**
4	**4**	**4.14**	**fractura**	**5**	**191**	**-1.12**	**justícia**
5	6	3.87	renovació	2	77	-1.13	cas
3	4	3.72	nord_de_el_país	**1**	**41**	**-1.22**	**europea**
3	4	3.72	socis	1	41	-1.22	part
3	**4**	**3.72**	**pacte**	**1**	**42**	**-1.26**	**la_justícia_espanyola**
3	4	3.72	informatius	1	42	-1.26	torra
2	3	3.55	municipis	**1**	**43**	**-1.29**	**lliurament**
2	**3**	**3.55**	**groller**	**1**	**44**	**-1.32**	**llei**
2	**3**	**3.55**	**fraudulent**	1	46	-1.39	pdecat
2	**3**	**3.55**	**nivell**	**2**	**99**	**-1.49**	**ordre**
2	**3**	**3.55**	**fonts_pròximes**	**1**	**51**	**-1.54**	**jutges**
2	**3**	**3.55**	**patriota**	**1**	**53**	**-1.59**	**consellers**
2	3	3.55	noms	**1**	**54**	**-1.62**	**fiança**
3	**5**	**3.40**	**tancat**	1	55	-1.65	partit
2	4	3.14	normal	**1**	**59**	**-1.75**	**defensa**
2	**4**	**3.14**	**els_serveis_jurídics**	**1**	**80**	**-2.19**	**independència**
2	4	3.14	petita	**2**	**198**	**-2.49**	**president**
2	4	3.14	optat	1	104	-2.56	belga
2	**4**	**3.14**	**policies**	1	147	-3.06	espanyol
3	**6**	**3.14**	**acompanyants**				
2	**4**	**3.14**	**presons**				
2	4	3.14	amic				
7	14	3.14	social				
2	**4**	**3.14**	**polítiques**				
2	4	3.14	milers				
3	**6**	**3.14**	**candidats**				
3	**112**	**-1.09**	**llibertat**				

Taula 10: *Taula elaborada amb els resultats de les notícies de* El Periódico *amb els valors de la mesura IM.*

1.3.2.3. EL PUNT AVUI

EL PUNT AVUI							
VALOR POSITIU				POSITIU NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	**3**	**4.04**	**dubtosos**	**1**	**27**	**-0.72**	**traïció**
3	**3**	**4.04**	**alertat**	**3**	**82**	**-0.74**	**alemany**
3	3	4.04	sànchez	1	29	-0.82	alta
3	3	4.04	romeva	**1**	**29**	**-0.82**	**actuació**
4	**4**	**4.04**	**bandes_armades**	**4**	**117**	**-0.83**	**detenció**
3	3	4.04	notificació	**1**	**30**	**-0.87**	**instància**
5	5	4.04	arrimadas	1	30	-0.87	actes
3	**3**	**4.04**	**consellers_de_la_independència**	2	62	-0.92	altres
4	**5**	**3.72**	**individus**	**1**	**31**	**-0.92**	**lamela**
3	**4**	**3.62**	**cap_a_bèlgica**	1	31	-0.92	ciutadans
3	4	3.62	mesa	**1**	**32**	**-0.96**	**civil**
3	4	3.62	rull	**1**	**35**	**-1.09**	**portaveu**
5	**7**	**3.55**	**rebels**	**5**	**176**	**-1.10**	**espanya**
2	3	3.45	chambre	**2**	**72**	**-1.13**	**exconsellers**
2	3	3.45	conseil	**1**	**37**	**-1.17**	**dret**
2	**3**	**3.45**	**circumstàncies_excepcionals**	**1**	**37**	**-1.17**	**eleccions**
2	**3**	**3.45**	**fonts_de_la_presó**	**3**	**114**	**-1.21**	**català**
2	**3**	**3.45**	**triomf**	1	38	-1.21	nou
2	**3**	**3.45**	**deriva**	**2**	**77**	**-1.23**	**brussel·les**
2	**3**	**3.45**	**propi_partit**	1	42	-1.36	cgpj
2	**3**	**3.45**	**els_serveis_secrets_espanyols**	**1**	**42**	**-1.36**	**presos**
2	**3**	**3.45**	**secrets**	**1**	**46**	**-1.49**	**causa**
2	3	3.45	desenes	**1**	**46**	**-1.49**	**independentistes**
4	**6**	**3.45**	**presó_preventiva**	**2**	**99**	**-1.59**	**ordre**
2	**3**	**3.45**	**querella_de_la_fiscalia**	**1**	**51**	**-1.64**	**jutges**
4	6	3.45	turull	**2**	**104**	**-1.66**	**belga**
2	**3**	**3.45**	**criminal**	**1**	**55**	**-1.74**	**política**
2	3	3.45	processament	**1**	**174**	**-3.41**	**expresident**
2	3	3.45	cambra				
3	**5**	**3.30**	**radicalisme**				
5	9	3.19	psc				
2	4	3.04	porta_tancada				
2	4	3.04	tancada				
2	4	3.04	opcions				
2	4	3.04	apel·lació				
2	**4**	**3.04**	**el_president_espanyol**				

Taula 11: *Taula elaborada amb els resultats de les notícies de* El Punt Avui *amb els valors de la mesura IM.*

1.3.2.4. VILAWEB

VILAWEB							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	NPT	PARAULA
3	**3**	**4.41**	**activar_el_consell_ de_la_república**	**1**	**46**	**1464**	**independentistes**
3	**3**	**4.41**	**internacionalitzant_ la_causa**	**1**	**46**	**1464**	**pdecat**
3	3	4.41	ben	**1**	**47**	**1464**	**república**
5	5	4.41	emmerson	**4**	**192**	**1464**	**presó**
3	3	4.41	onu	**1**	**49**	**1464**	**polític**
3	**3**	**4.41**	**ministeri_d'_afers_estrangers**	**1**	**51**	**1464**	**catalans**
3	**3**	**4.41**	**testimoni_de_el_naixement**	**1**	**52**	**1464**	**públics**
5	**5**	**4.41**	**naixement**	**3**	**176**	**1464**	**espanya**
5	**6**	**4.15**	**nació**	**1**	**61**	**1464**	**demanda**
4	5	4.09	recordat	**1**	**72**	**1464**	**violència**
10	**13**	**4.03**	**retorn**	**1**	**82**	**1464**	**alemany**
3	4	4.00	conferència_de_premsa	**1**	**96**	**1464**	**procés**
3	**4**	**4.00**	**testimoni**				
2	**3**	**3.83**	**tribunal_d'_apel·lació**				
2	**3**	**3.83**	**decisió_de_la_justícia**				
4	**6**	**3.83**	**els_presos_polítics**				
2	**3**	**3.83**	**exigit**				
2	**3**	**3.83**	**reiterat**				
2	3	3.83	argumentat				
2	3	3.83	centenars				
2	3	3.83	familiars				
2	3	3.83	posterior				
2	**3**	**3.83**	**desacreditada**				
2	**3**	**3.83**	**alliberament_de_els_presos**				
2	**3**	**3.83**	**impacte**				
3	**5**	**3.67**	**periple**				
4	**7**	**3.60**	**repressió**				
5	**9**	**3.56**	**casa_de_la_república**				

Taula 12: *Taula elaborada amb els resultats de les notícies de* Vilaweb *amb els valors de la mesura IM.*

1.3.3. Diaris originalment en anglès

1.3.3.1. INDEPENDENT

INDEPENTENT							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
11	**11**	**5.45**	**sr._puigdemont**	**1**	**44**	**-0.01**	**llei**
4	**4**	**5.45**	**declarar_la_independència**	**1**	**47**	**-0.10**	**independentista**
5	**6**	**5.19**	**fiscals**	**1**	**48**	**-0.14**	**tribunal_suprem**
3	4	5.03	gabinet	**4**	**192**	**-0.14**	**presó**
2	3	4.86	predecessor	**2**	**102**	**-0.22**	**espanyola**
2	3	4.86	estatal	**1**	**51**	**-0.22**	**jutges**
2	**3**	**4.86**	**ordres_d'_arrest**	**1**	**52**	**-0.25**	**detingut**
2	**3**	**4.86**	**autoimposat**	**1**	**57**	**-0.38**	**delictes**
2	**3**	**4.86**	**l'_ex_líder**	**1**	**77**	**-0.82**	**brussel·les**
2	**3**	**4.86**	**espera_d'_una_decisió**	1	77	-0.82	cas
2	**3**	**4.86**	**ús_de_els_fons**	**2**	**163**	**-0.90**	**bèlgica**
3	**6**	**4.45**	**càrrecs_de_rebel·lió**	**1**	**84**	**-0.94**	**malversació**
2	**4**	**4.45**	**pagar_una_fiança**	1	91	-1.06	dies
2	**4**	**4.45**	**president_de_catalunya**	**1**	**128**	**-1.55**	**llarena**
2	5	4.13	regió	**1**	**140**	**-1.68**	**estat**
6	**16**	**4.03**	**ex**	**4**	**634**	**-1.86**	**puigdemont**
1	3	3.86	dur	**1**	**223**	**-2.35**	**jutge**
2	6	3.86	antic				
1	**3**	**3.86**	**forma_de_república**				
1	3	3.86	cambra				
1	**3**	**3.86**	**acomiadat**				
1	3	3.86	desenes				
1	3	3.86	xarxa				
1	**3**	**3.86**	**un_tribunal_alemany**				
1	**3**	**3.86**	**rebel**				
1	**3**	**3.86**	**fugitius**				
1	3	3.86	desenvolupament				
1	**3**	**3.86**	**triomf**				
1	**3**	**3.86**	**abandonat**				
1	3	3.86	esforços				
1	**3**	**3.86**	**el_govern_regional**				
1	3	3.86	pendent				
1	**3**	**3.86**	**espanya_per_càrrecs**				
1	**3**	**3.86**	**el_polític_independentista**				

Taula 13: *Taula elaborada amb els resultats de les notícies de Independent amb els valors de la mesura IM.*

1.3.3.2. THE GUARDIAN

THE GUARDIAN							
VALOR POSITIU				VALOR NEGATIU			
APD	APT	IM	PARAULA	APD	APT	IM	PARAULA
3	**3**	**4.29**	**declaració_ d'_independència**	1	41	-1.07	part
3	**3**	**4.29**	**càrrega**	**1**	**42**	**-1.11**	**la_justícia_espanyola**
3	3	4.29	col·legues	2	89	-1.19	país
3	**3**	**4.29**	**ministre_espanyol**	2	91	-1.22	dies
3	**3**	**4.29**	**elegit_president**	**1**	**46**	**-1.24**	**independentistes**
3	**3**	**4.29**	**un_nou_govern**	**1**	**46**	**-1.24**	**causa**
4	**5**	**3.97**	**els_fons_públics**	**2**	**96**	**-1.30**	**procés**
6	**8**	**3.87**	**indegut**	**1**	**51**	**-1.39**	**jutges**
3	**4**	**3.87**	**l'_ús_indegut**	**1**	**54**	**-1.47**	**fiança**
5	**7**	**3.80**	**elegit**	**3**	**174**	**-1.57**	**expresident**
2	**3**	**3.70**	**caiguda**	**1**	**59**	**-1.60**	**defensa**
2	**3**	**3.70**	**espanya_per_càrrecs**	**1**	**61**	**-1.64**	**petició**
2	**3**	**3.70**	**el_nou_govern**	**1**	**61**	**-1.64**	**demanda**
2	3	3.70	converses	**2**	**128**	**-1.71**	**llarena**
2	3	3.70	regionals	**3**	**192**	**-1.71**	**presó**
2	3	3.70	realitat	1	65	-1.74	fet
2	**3**	**3.70**	**un_tribunal_alemany**	**1**	**68**	**-1.80**	**judicial**
2	**3**	**3.70**	**ús_de_fons**	**2**	**140**	**-1.84**	**estat**
5	**8**	**3.61**	**unilateral**	**1**	**73**	**-1.90**	**advocat**
3	**5**	**3.55**	**exiliat**	1	77	-1.98	cas
3	**5**	**3.55**	**detenció_internacional**	**1**	**84**	**-2.11**	**malversació**
3	5	3.55	proper	**2**	**223**	**-2.51**	**jutge**
3	5	3.55	simple	**1**	**112**	**-2.52**	**llibertat**
6	11	3.41	directa	**1**	**125**	**-2.68**	**delicte**
2	**4**	**3.29**	**garantia**	**1**	**191**	**-3.29**	**justícia**
2	4	3.29	llenguatge				
2	4	3.29	normalitat				
2	4	3.29	inés_arrimadas				
2	**4**	**3.29**	**presons**				
2	**4**	**3.29**	**objectius**				
11	23	3.22	ús				
3	**7**	**3.06**	**mandat**				
3	**7**	**3.06**	**més_greu**				
11	**26**	**3.05**	**catalana**				
5	**12**	**3.02**	**ministre**				

Taula 14: *Taula elaborada amb els resultats de les notícies de* The Guardian *amb els valors de la mesura IM.*