UNIVERSITAT DE BARCELONA

Grau de Filologia Catalana

Treball de Fi de Grau

Curs 2023-2024

L'afàsia i el processament del llenguatge natural: anàlisi computacional d'un corpus amb spaCy

Núria Poch Franco

Tutora: Mireia Farrús Cabeceran

Barcelona, 18 de juny de 2024

AGRAÏMENTS

A la Mireia, la meva tutora, per haver-me escoltat i aconsellat en tot el desenvolupament del treball, i per haver-me transmès la passió per la lingüística computacional. Als investigadors del projecte de la Universitat Politècnica de Catalunya: al Toni Hernández i a la Neus Català, per acollir-me des del primer moment i, especialment, a la Maite Zaragoza, per acompanyar-me i recolzar-me sempre. I, en darrer terme, però no menys important, a la meva família i amics, per animar-me a no parar mai d'aprendre.

RESUM

Avaluar adequadament la parla de les persones amb afàsia afavoreix en la recuperació de la seva capacitat comunicativa. A conseqüència de l'evolució tecnològica actual, és possible l'automatització de tasques bàsiques d'aquest procés. En aquest context, un projecte de la Universitat Politècnica de Catalunya s'ha proposat crear un sistema que detecti automàticament el grau de gravetat de l'afàsia en català. El treball següent s'emmarca dins d'aquest projecte amb un doble objectiu: participar-hi en la part de preprocessament de les dades, amb la transcripció de discursos amb diversos tipus d'afàsia. I, a partir d'aquests, crear i executar un algorisme propi que permeti l'automatització de la seva anàlisi per mitjà de spaCy, una llibreria de Python. Durant el treball s'ha pogut constatar que les eines actuals de processament del llenguatge natural en català presenten certes limitacions per executar anàlisis completes, a causa del reduït nombre de recursos disponibles i per la manca de dades anotades en aquesta llengua.

Paraules clau: *afàsia*, *processament del llenguatge natural*, *spaCy*, *lingüística computacional*, *català*.

ABSTRACT

Assessing properly the speech of people with aphasia favours the recovery of their communicative abilities. As a result of current technological evolution, it is possible the automation of basic tasks in the process. In this context, a project ruled by the Polytechnic University of Catalonia aims to create a system that automatically detects the severity degree of aphasia in Catalan language. The work presented here is framed within that project with a double objective: to participate in the data preprocessing phase, involving the transcription of speeches with various types of aphasia, and to develop and execute an algorithm that enables the automation of their analysis using spaCy, a Python library. During the course of this work, it has been observed that current natural language processing tools for Catalan have certain limitations in performing complete analyses, due to the limited number of available resources and the lack of written data in that language.

Key words: *aphasia*, *natural language processing*, *spaCy*, *computational linguistics*, *Catalan*.

UNIVERSITAT DE BARCELONA

Coordinació d'Estudis
Facultat de Filologia

Gran Via
de les Corts Catalanes, 585
08007 Barcelona

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

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

Barcelona, a 18 de juny de 2024

Signatura:

Núria Poch Franco

ÍNDEX

1. INTRODUCCIÓ

Les lesions al cervell poden afectar de manera directa a diversos factors de la vida d'una persona, com per exemple en alteracions al moviment o en trastorns de la parla. Un accident cerebrovascular, un traumatisme o un tumor cerebral poden produir afàsia, que és definida per Faustino Diéguez i Jordi Peña com l'alteració de la conducta verbal d'una persona a causa d'un dany en el cervell (DIÉGUEZ i PEÑA, 2012: 4). Un discurs produït per una persona amb afàsia pot estar format per característiques i nivells de gravetat diferents, fet que dependrà de la zona del cervell on hagi succeït la interrupció del flux sanguini o a on es localitzi el tumor o traumatisme, i també del grau d'afectació. Per això les afàsies s'han classificat en diversos tipus, segons les seves característiques. Per una banda, en les anomenades afàsies fluents es manté la capacitat de producció oral, però es veu limitada la de comprensió, dins de les quals s'inclouen l'afàsia de Wernicke, la de conducció, la transcortical sensorial i l'anòmica. I, per altra banda, les afàsies no fluents, en què es conserva la capacitat de comprensió, però es veu alterada la de producció, i es donen afectacions en la parla com parafàsies, trastorns de l'articulació verbal o estereotípies verbals, entre d'altres. Dins d'aquest segon grup hi ha l'afàsia transcortical mixta i la motora, la global i la de Broca.

L'avaluació de les capacitats lingüístiques d'una persona amb afàsia és una tasca imprescindible per al seu tractament, atès que ajuda a diagnosticar-ne el tipus, la gravetat i si hi ha trastorns associats, i a dissenyar un programa d'intervenció i rehabilitació el més adaptat i individualitzat a cada cas (GÓMEZ, 2013: 69). Existeix una elevada variabilitat de característiques lingüístiques a examinar, com per exemple la fluïdesa verbal, que es relaciona amb el nombre de paraules i la longitud de les oracions, l'entonació o el contingut informatiu (2013: 71). Aquestes característiques es poden recollir manualment per part del professional sanitari o terapeuta, però gràcies a l'evolució de les tecnologies actuals i de la intel·ligència artificial, cada cop són més les tasques que es poden automatitzar, i així aconseguir un procés d'avaluació més ràpid. El processament del llenguatge natural (PLN) és una disciplina clau en aquest àmbit, atès que és la branca de la intel·ligència artificial que s'encarrega que el llenguatge humà sigui comprensible pels ordinadors i sistemes informàtics.

En aquest context, la Universitat Politècnica de Catalunya (UPC), juntament amb l'Institut d'Estudis Catalans (IEC), van engegar l'any 2022 el projecte que porta per nom «Semàntica

de les paraules del català: teoria i aplicacions clíniques»[1], en el qual s'està desenvolupant un estudi estadístic de la semàntica i la sintaxi del català, relacionat amb l'àmbit clínic i els trastorns del llenguatge. Així, per mitjà de la comparació de la producció lingüística de persones sense patologies de la parla amb persones amb quadres d'afàsia, es vol aconseguir un model de reconeixement de la parla i del grau de gravetat o quocient d'afàsia en parlants afàsics en català.

L'origen del treball final de grau que es presenta a continuació sorgeix de la participació en aquest projecte de la UPC i l'IEC. Durant el treball es durà a terme una doble tasca: per una banda, es formarà part de l'equip investigador del projecte, dins el subgrup de clínica, on es treballarà en la part de preprocessament de les dades amb la transcripció dels àudios enregistrats dels participants amb afàsia al projecte. I, per altra banda, a partir d'aquestes dades i transcripcions, s'intentarà automatitzar els indicadors d'anàlisi discursiva establerts per Michel Paradis al seu *Test d'afàsia per a bilingües*, per mitjà de la creació d'un algorisme propi amb una eina de processament del llenguatge natural en català. Després, s'analitzaran algunes de les transcripcions del projecte amb diversos tipus d'afàsia amb l'ús de l'algorisme creat. A partir d'aquestes tasques, l'autora del treball es planteja dues preguntes de recerca: quines eines de PLN es poden usar actualment per automatitzar l'anàlisi lingüística en català? I què permeten fer aquestes eines i quines són les seves limitacions en català?

Així doncs, l'objectiu general d'aquest treball és determinar l'eficàcia d'una eina de PLN per analitzar la parla d'un corpus de persones afàsiques en català. Es parteix de dos objectius específics: primerament, desenvolupar un algorisme propi amb una eina de PLN que permeti automatitzar indicadors discursius per a l'anàlisi de la parla afàsica. I, després, formar part i conèixer el funcionament d'un projecte de recerca amb dades clíniques, i participar en el procés de transcripció i preprocessament de les dades. Partint d'aquests objectius, la hipòtesi que es formula és que amb les eines de PLN disponibles actualment és possible fer una anàlisi lingüística en català limitada, per la manca d'entrenament i de dades anotades, i per la reduïda disponibilitat de recursos en aquesta llengua.

Per assolir aquests objectius, tot seguit s'explicarà l'estructura que seguirà aquest treball. En primer lloc, es definirà l'afàsia i els diversos tipus que existeixen, amb les característiques

[1] Vegeu pàgina web del projecte: https://futur.upc.edu/32847767

lingüístiques corresponents de cada un. Es parlarà sobre la importància d'analitzar la parla de les persones amb aquest trastorn del llenguatge, i sobre què ha d'incloure una avaluació perquè sigui adequada i quins tipus existeixen, com els tests o les bateries. Un cop entès el concepte d'afàsia, s'explicarà el funcionament del projecte de la UPC i l'IEC, els seus objectius i totes les accions que s'estan duent a terme. I es tractarà concretament de la tasca desenvolupada per l'autora del treball dins d'aquest projecte, amb la transcripció dels àudios dels participants en l'estudi, i totes les problemàtiques i qüestions sorgides durant la transcripció. Després, el treball se centrarà en la cerca d'eines de PLN, focalitzada en les llibreries que hi ha actualment obertes, gratuïtes i amb funcionalitats disponibles en català a Internet. S'escollirà una llibreria en funció de les necessitats del treball i es durà a terme un procés d'aprenentatge autònom de la seva arquitectura i del procés d'elaboració dels codis. Amb aquests coneixements, es crearà un algorisme per a la posterior execució i anàlisi de les transcripcions elegides. S'explicarà de manera detallada aquest algorisme, amb el guió d'instruccions i els diferents passos que el formen, i algunes de les problemàtiques sorgides durant la seva generació. I, finalment, es presentaran els resultats obtinguts: per una banda, sobre el funcionament de la llibreria i de les limitacions que s'han trobat pel cas del català. I, per l'altra, es faran observacions sobre els resultats dels discursos analitzats, en funció de les característiques lingüístiques del tipus d'afàsia de cada persona. En darrer terme, es farà un recull de les conclusions més importants de tot el procés de realització del treball.

2. MARC TEÒRIC

2.1. *L'afàsia i les característiques lingüístiques de cada tipus*

L'afàsia es defineix com l'alteració de la conducta verbal d'una persona a causa d'una lesió cerebral (DIÉGUEZ i PEÑA, 2012: 4), en la qual poden veure's afectats processos responsables de la interpretació, producció i formulació dels símbols del llenguatge, ja sigui un fonema, una síl·laba, un sintagma, una oració o un discurs (AGUADO et al., 2013: 38). La lesió cerebral pot tenir etiologies diverses, com un tumor o un traumatisme en aquesta zona, tot i que la causa més freqüent són els accidents cerebrovasculars, en els quals hi ha una interrupció del flux sanguini a una determinada localització cerebral relacionada amb el llenguatge (UGGEN, 2020: 1). Aquest dany afecta la capacitat de parlar i es poden donar alteracions en la comprensió i producció del llenguatge oral, la lectura o l'escriptura (WEBB i ADLER, 2010: 237), i dependrà del tipus i del grau de gravetat (pot anar de lleu a molt greu).

Existeixen diverses classificacions, en funció de la zona cerebral i de les capacitats afectades. Una de les distribucions més emprades és la distinció en dos grans grups: les afàsies fluents i les no fluents. Les afàsies fluents són aquelles en les quals la persona és capaç d'emetre un discurs gramaticalment i prosòdic correcte, però la capacitat de comprensió es veu reduïda. Altrament, la seva parla pot contenir neologismes, un argot propi o un ús inadequat de les paraules. Per altra banda, en les afàsies no fluents la comprensió es manté normalment sense afectacions, però hi ha certes dificultats a l'hora de produir el discurs, amb construccions gramaticalment incorrectes o segments inintel·ligibles (UGGEN, 2020: 2). Els símptomes més greus d'aquest segon tipus són el mutisme, les vocalitzacions o les estereotípies (DIÉGUEZ i PEÑA, 2012: 82).

Una altra característica que es pot observar per diferenciar els tipus d'afàsia és si els individus conserven la capacitat de repetició o si la tenen alterada. Els tipus que presenten aquesta mancança són l'afàsia de Broca, la de Wernicke, la de conducció i la global, i els que la tenen preservada són l'afàsia transcortical motora, la mixta, la sensorial i l'anòmica (2012: 83). Aquesta classificació de les afàsies prové de l'Escola de Boston, que van crear el *Test de Boston per a l'avaluació de les afàsies*, considerat com un desenvolupament de la classificació proposada per Carl Wernicke i Ludwig Lichtheim (2012: 82). El model de Wernicke-Lichtheim associa les capacitats de comunicació amb diferents regions cerebrals, i diferencia tres tipus d'afàsies: la de Broca, la de Wernicke i l'anòmica (TORRE et al., 2021: 34). Tot seguit, es presenta la imatge d'un cervell, en la qual hi ha representades les diverses zones de cada tipus d'afàsia.

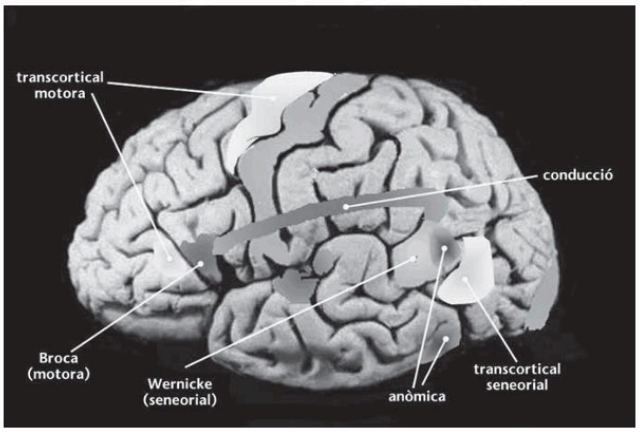


Fig. 1. Àrees del cervell (AGUADO et al., 2013: 41)

Per altra banda, a continuació es presenta una taula de les característiques lingüístiques de cada tipus d'afàsia, pel que fa a l'expressió oral[2].

TIPUS D'AFÀSIA	EXPRESSIÓ ORAL	COMPRENSIÓ	REPETICIÓ	DENOMI-NACIÓ
Broca (o motora)	No fluent. A l'inici pot haver-hi mutisme, vocalitzacions inarticulades i estereotípies verbals. L'evolució de l'afàsia pot donar anàrtria, anòmia o agramatisme[3]. També hi ha una reducció i simplificació de les oracions, fins al punt que pot arribar a una parla telegràfica, amb l'ús de paraules i expressions d'ús freqüent. Hi ha una conservació relativa de l'ús dels substantius i, en menor grau, dels verbs i adjectius (JUNQUET i BARROSO, 1994: 274).	Bona en general. Pot haver-hi algun problema de comprensió en ordres complexes i textos, o en oracions amb clàusules subordinades.	Alterada.	Alterada. Millora amb ajudes o claus fonèmiques.
Wernicke	Fluent, amb una articulació i prosòdia normal. Però amb un excés de producció verbal, ús d'argot i neologismes o parafàsies fonològiques propis de l'individu (DIÉGUEZ i PEÑA, 2012: 90). En les primeres fases, no són conscients que no se'ls entén, fenomen conegut com a anosognòsia (AGUADO et al., 2013: 42).	Greument afectada a tots els nivells (paraules, frases, discursos, etc.). Les dificultats adopten diferents formes segons si la lesió està més o menys a prop d'altres nuclis neuronals (2013: 42).	Alterada, per les dificultats per discriminar els components de la seqüència (2013:42).	Alterada.

[2] De l'escrita no se'n parlarà, per raons d'extensió del treball, i també pel fet que l'estudi del projecte de la UPC s'està analitzant només l'expressió oral.
[3] Vegeu *Annex II*, en el qual hi ha un glossari de paraules clau en la terminologia de les afàsies.

Conducció	Fluent. El seu trastorn principal se centra en la repetició verbal, amb presència de parafàsies fonèmiques (DIÉGUEZ i PEÑA, 2012: 94).	Bona.	Alterada, tret més característic d'aquest tipus. No poden traslladar la informació escoltada a l'àrea encarregada de programar la producció perquè les vies que uneixen ambdues àrees estan lesionades (AGUADO et al., 2013: 43).	Alterada. Millora amb ajudes.
Global	No fluent. Expressió molt reduïda, que pot arribar al mutisme i amb alguna vocalització del tipus *ah* o *eh*, o esterotípies síl·làbiques (2012: 98). Hi ha una possible preservació del llenguatge automàtic. És el tipus d'afàsia que usualment dona presentacions més greus.	Molt alterada.	Preservada.	Alterada.
Motora transcorti-cal	No fluent. Mutisme o reducció de la parla a l'inici. Té uns símptomes semblants a l'afàsia de Broca, però amb una bona repetició (AGUADO et al., 2013: 39). Per això, es poden donar algunes	Bona.	Preservada.	Alterada, tot i que en alguns casos preservada.

	perseveracions i ecolàlies en resposta a preguntes, i problemes d'accés al lèxic que provoquen intents i latències, i respostes del tipus "no ho sé" o "no puc" (DIÉGUEZ i PEÑA, 2012: 103). Durant l'evolució, sol millorar la fluència i apareixen frases breus (2012: 101).			
Sensorial transcorti-cal	Fluent, amb possibles ecolàlies.	Afectada.	Preservada.	Alterada.
Mixta transcorti-cal	No fluent. Limitada a repeticions automàtiques (ecolàlies) i reduïda a produccions inarticulades i intel·ligibles, estereotípies i alguna expressió incompleta (DIÉGUEZ i PEÑA, 2012: 112).	Molt afectada.	Preservada.	Alterada.
Anòmica	Fluent i amb una bona estructura gramatical. La característica més destacable és l'anòmia, és a dir, un dèficit en l'evocació de paraules, que provoca latències en la selecció de paraules (2012: 113). També possibles circumloquis i paraules òmnibus (del tipus *això, cosa*, etc.).	Bona.	Preservada.	Alterada.

2.2. *L'anàlisi discursiva de persones amb afàsia*

L'afasiologia és la disciplina que s'encarrega d'estudiar les afàsies, les lesions neurològiques que la produeixen i les conductes verbals relacionades (DIÉGUEZ i PEÑA, 2012: 4). Les capacitats lingüístiques són un dels indicadors més evidents per detectar quin grau i tipus d'afàsia té la persona. Per aconseguir una bona avaluació d'aquestes, són necessaris instruments de mesura objectius i estandarditzats que analitzin adequadament els dèficits i la gravetat de la patologia (GÓMEZ, 2013: 109). Un examen que analitzi l'afàsia ha d'incloure tasques verbals per avaluar l'expressió oral, la repetició, la denominació, l'evocació de les paraules, la comprensió verbal, la lectura i l'escriptura (1993: 69). Normalment, es desenvolupa per mitjà d'una conversa entre la persona i un professional sanitari, en la qual aquest li demana certes tasques que impliquen la producció verbal, ja sigui amb la descripció d'una imatge o amb l'explicació d'esdeveniments quotidians de la seva vida, de manera que es valora l'expressió espontània sobre temes diversos (WRIGHT, 2011: 1283). Nicole Müller, al llibre *The Handbook of Clinical Linguistics*, fa una distinció dels diversos tipus de discursos que es troben en l'àmbit clínic. Dins de l'avaluació de les afàsies es destaca el tipus descriptiu, que és el que es produeix en el moment de la descripció d'una imatge, i on la persona produeix una llista d'atributs i conceptes amb el propòsit de traduir la imatge a llenguatge (BALL, 2008: 23). També hi ha el tipus narratiu, quan per exemple s'explica una experiència personal a través d'esdeveniments i, en darrer terme, el tipus procedimental, quan s'ofereixen instruccions o direccions.

Entre les diverses avaluacions per estudiar l'afàsia es poden distingir els tests i les bateries. Per una banda, els tests tenen com a objectiu estudiar la presència o absència d'alteracions afàsiques (GÓMEZ, 2013: 74). D'exemples de testos hi ha el *Test Barcelona* o el *Test de Boston per al diagnòstic de l'afàsia*, que se centren en les característiques lingüístiques produïdes a partir de l'execució d'activitats com la descripció d'una imatge (MIR, 2022: 5). Per altra banda, les bateries d'afàsia no només identifiquen les alteracions, sinó que també permeten conèixer la gravetat i proporcionen dades sobre altres aspectes de la funcionalitat lingüística, a partir de la puntuació que obté cada persona en els diferents subtests. Com a exemples de bateries destaquen el *Test d'Afàsia per a Bilingües*, de Michel Paradis i Gary Libben, o la *Bateria d'Afàsies Western*.

Tanmateix, l'heterogeneïtat d'aquestes avaluacions pot presentar alguns problemes, com la necessitat d'adaptar-se a les característiques específiques d'una llengua o l'elevada quantitat de paràmetres que es poden observar (UGGEN, 2020: 3). A més, com s'ha pogut constatar anteriorment, la producció discursiva de les persones amb afàsia és molt diversa. Així, els afàsics tipus Broca tenen tendència a produir discursos breus, amb pocs verbs i quasi cap estructura sintàctica complexa. En canvi, els tipus Wernicke produeixen un discurs molt més extens, però sovint incoherent (DIÉGUEZ i PEÑA, 2012: 193). Tot i aquestes diferències, Diéguez recalca el fet que sí que existeixen algunes característiques comunes, com l'ús reduït de marcadors cohesius i connectors, la utilització d'un gran nombre de díctics o de pronoms de tercera persona (2012: 193). Altrament, afirma que els individus afàsics presenten una preservació de la superestructura, definida com l'organització del discurs, i una afectació de la macroestructura, que seria el contingut general d'aquest (2012: 289).

Un altre aspecte crucial a tenir en compte a l'hora d'analitzar el discurs en persones amb afàsia és el bilingüisme, atès que és un context que presenta unes característiques concretes i requereix instruments i mètodes diferenciats (DIÉGUEZ i PEÑA, 2012: 119). Michel Paradis, lingüista canadenc, va fer un pas important per a l'avaluació de persones afàsiques bilingües amb la creació del *Test d'afàsia per a bilingües*, disponible i adaptat en més de seixanta llengües. L'inici de la descripció de l'afàsia es remunta a la segona meitat del segle XIX, on gran part dels estudis es van fer en societats monolingües (GÓMEZ, 2013: 109), i per aquest motiu va ser bàsica la tasca de Paradis. La finalitat del test és determinar si l'actuació lingüística en una de les llengües és millor, pitjor o igual que en l'altre, i també ajuda a determinar quina és la més accessible per dur a terme el tractament. Les diferents versions d'aquest test no són traduccions d'una llengua a l'altre, sinó que són adaptacions a cada una en funció de les seves característiques lingüístiques i culturals. Les versions d'aquest test pel castellà i el català van ser elaborades per Josep Elias (2013: 110).

2.3. *Projecte de recerca: Semàntica de les paraules en català: teoria i aplicacions clíniques*

El títol del projecte de la UPC i l'IEC *Semàntica de les paraules del català: teoria i aplicacions clíniques* defineix el seu objectiu principal: fer un estudi estadístic de la semàntica i la sintaxi del català en l'àmbit clínic, a partir de la comparació de la producció

lingüística amb persones sense trastorns de la parla, respecte a persones que han patit quadres clínics d'afàsia. D'aquesta manera, s'estén l'estudi estadístic del català al cas de les patologies del llenguatge, amb l'objectiu d'aconseguir un model de reconeixement de la parla en catalanoparlants afàsics, que permeti la transcripció automàtica dels seus discursos i també la detecció del quocient o grau de gravetat de l'afàsia. Dins del projecte s'han establert dos subgrups: el subgrup de recerca clínica, que s'encarrega de la recopilació i anàlisi de les dades clíniques, i el subgrup de computacional, que se centra en el desenvolupament de programes informàtics i de models estadístics a partir de les dades clíniques.

Primerament, convé tenir present que com que les dades amb què es treballa en aquest projecte són clíniques i confidencials, esdevé imprescindible l'aprovació per part d'un comitè d'ètica. Així, els protocols van ser establerts i aprovats pel Comitè d'Ètica per a la Investigació Clínica del Consorci Sanitari Integral (CSI) i pel Comitè d'Ètica de la UPC. Un cop aprovats els permisos, es va iniciar el reclutament de pacients el novembre del 2022. Els participants provenen de dos centres sanitaris: l'Hospital General de l'Hospitalet i de la Clínica de Logopèdia FLA-Universitat de València. En data de març de 2024, s'ha aconseguit la participació de vuit persones a l'Hospitalet, sis dels quals eren bilingües de català i castellà i dos monolingües castellans. I a València han participat onze persones, cinc bilingües i sis monolingües. Per tant, es disposa d'un total d'onze individus bilingües i de sis monolingües. L'etiologia de l'ictus fou isquèmica en quatre casos i hemorràgica la resta. I els tipus d'afàsia són nou persones amb afàsia motora o de Broca, tres afàsies globals, tres afàsies anòmiques i dues afàsies transcorticals motores.

Els procediments realitzats que s'han dut a terme per part del subgrup de clínica han estat la recopilació de persones, la recollida del consentiment informat, de les dades sociodemogràfiques (edat, data, lloc de naixement, nivell d'estudis i anys d'escolarització) i de les clíniques (data d'ingrés hospitalari, tipografia lesional, tipus d'afàsia, data d'alta hospitalària i informació sobre la realització o no de sessions de logopèdia ambulatòria). També s'ha analitzat l'historial de bilingüisme a partir del qüestionari que inclou el *Test d'Afàsia per a Bilingües* de Paradis i Libben. Finalment, totes les dades recollides s'han inclòs en un paquet d'anàlisi estadístic SPSS (*Statistical Package for Social Sciences*), per a un estudi estadístic descriptiu. Per a realitzar l'estudi amb els participants s'ha emprat l'adaptació al català i castellà de la *Bateria d'Afàsies Western* (BAW), que Maite Zaragoza i el seu equip va realitzar dins d'aquest mateix projecte (ZARAGOZA, 2022). Després, en data

de maig de 2024 s'està duent a terme el preprocessament de les dades. S'han enregistrat totes les respostes de parla espontània i descriptiva en format d'àudio, que estan sent transcrites. Totes les transcripcions seran enviades al subgrup de computacional, que tindrà com a activitat principal la modelització matemàtica i estadística del català, i l'obtenció d'un model de reconeixement de veu a partir de la base de dades clíniques recollida. Així, gràcies a aquesta tecnologia, en un futur es podran crear aplicacions que ajudin al procés de recuperació i rehabilitació de trastorns del llenguatge com l'afàsia, o a aportar nous coneixements per a la seva investigació.

3. MARC EXPERIMENTAL

3.1. *El procés de transcripció de discursos afàsics*

Tot seguit s'explicarà el procés de transcripció, que és la tasca que s'ha fet com a participació en el projecte de la UPC. Les dades amb les quals s'ha treballat són enregistraments de persones amb afàsia durant una sessió de teràpia. Per a poder dur a terme el tractament de les dades per part del subgrup de computacional, és necessària una fase prèvia que porta per nom preprocessament de les dades. Són diverses les accions que es poden fer en aquesta etapa i dependrà de les necessitats de cada projecte. Algunes d'aquestes són la transcripció d'un àudio de veu a text, l'estandardització a majúscules o minúscules, l'eliminació d'accents, d'espais en blanc o de determinades seqüències que no interessen, o la detecció d'una determinada seqüència de caràcters o *strings* (VEGAS, 2010: 16). Un preprocessament adequat és essencial perquè l'anàlisi i els resultats siguin correctes. En aquest projecte en concret el preprocessament de les dades s'ha basat a tenir els àudios transcrits a partir d'unes convencions marcades i adaptades a les seves necessitats i característiques.

L'autora del treball ha transcrit un total de 28 minuts i 45 segons d'enregistraments en català, i un total de 2 hores, 47 minuts i 32 segons en castellà, per mitjà de l'aplicació Buzz, que s'explicarà tot seguit. Com es pot comprovar, hi ha una diferència evident entre la quantitat de minuts entre castellà i català. Per cada persona hi ha dos tipus d'àudio: un primer en què es desenvolupa una parla espontània entre el professional i la persona, i un segon en el qual aquesta descriu un dibuix. Els enregistraments s'han guardat amb títols que no contenen el seu nom, sinó que estan formats per nombres: 001 si l'àudio prové de València i 002 si prové de l'Hospitalet de Llobregat. Després, cada individu està identificat amb un número i es

diferencia si és català o castellà (CAT o CAST o ESP). Un exemple de títol seria 001.003.CAT.

El preprocessament de les dades és una de les parts que requereix més temps, especialment si es fa una transcripció manual, i per això normalment s'usen programes informàtics que automatitzen aquest procés. Tanmateix, és només un suport, atès que aquests programes desconeixen les convencions que s'empren en cada projecte i és necessària una revisió manual. Per a dur a terme la transcripció en aquest projecte, s'ha fet ús del programa Buzz, que és un sistema de reconeixement automàtic de veu, el qual permet transcriure àudios i traduir en temps real. Funciona a través del model Whisper, que és un sistema d'intel·ligència artificial d'OpenAI que serveix específicament per transcriure arxius d'àudio a text (FERNÁNDEZ, 2023). En ser un sistema que treballa amb un model d'aprenentatge automàtic, succeeix que amb llengües com el castellà o l'anglès funciona adequadament, atès que se les ha pogut entrenar amb una elevada quantitat de dades. De fet, es considera que en castellà té una taxa d'error de menys del 5% (FERNÁNDEZ, 2023). En canvi, amb llengües més minoritàries com el català, ocasiona que el sistema cometi més errors, tal com s'ha pogut comprovar a l'hora de dur a terme les transcripcions. Una altra circumstància que ha succeït és que alguns dels àudios no se senten nítidament i s'ha fet ús del programa Audacity, un programari d'edició d'àudio gratuït, que permet amplificar-ne el so.

S'ha seguit el protocol de transcripció del projecte[4], que segueix les convencions que es presenten a continuació. Sota el terme **pausa plena** s'han agrupat les situacions següents: si la persona fa ús de crosses o falques del discurs (*filler words*), o sons del tipus "eh", "um" o "ah", o també si riu o emet interjeccions. Després, si les paraules o oracions no s'entenen s'ha escrit **inintel·ligible** i s'han marcat les ecolàlies en els moments que la persona repeteix la paraula o sintagma que li diu el professional. Les intervencions d'aquest no s'han transcrit i s'han marcat com a **interlocutor**. Si la persona comet errors de pronunciació o parafàsies al parlar, s'ha escrit la paraula objectiu corregida. L'ús de les majúscules i les minúscules s'ha conservat, s'han escrit tots els signes de puntuació, i la segmentació per oracions ha seguit els principis sintàctics i prosòdics del català i del castellà. No s'han plasmat els allargaments vocàlics ni consonàntics, ni tampoc s'ha tingut en compte la llargada de les pauses ni

[4] Vegeu *Annex II*.

comptabilitzat la durada de segons. En darrer terme, s'han preservat les repeticions i els desajustaments sintàctics, per mantenir els patrons de parla reals de les persones afàsiques.

La principal problemàtica que ha sorgit durant el procés de transcripció, pel que fa a les transcripcions de català, ha estat que ha sigut necessari refer quasi tot el text que executava l'aplicació Buzz, atès que confonia el català amb altres llengües i el resultat era una barreja de llengües incoherent. Fins i tot en alguns casos apareixien símbols de llengües orientals, com el japonès. Les mancances en detectar el català provoca, a més, un altre problema: ha segmentat les oracions erròniament, acció que no és modificable, ja que Buzz t'executa el discurs amb uns segments predeterminats en franges de segons. Això encara fa més complexa la transcripció, perquè hi ha hagut dificultats per encaixar les oracions i les pauses, característica essencial en un discurs afàsic. Per altra banda, les transcripcions en castellà han estat molt més eficaces i ràpides, pel fet que el sistema té una capacitat més bona de reconeixement en aquesta llengua i no ha estat necessari fer tants canvis. A més, convé destacar que l'ha detectat adequadament en la majoria dels casos, i no l'ha confós amb altres, i la segmentació en oracions no ha estat tan equívoca.

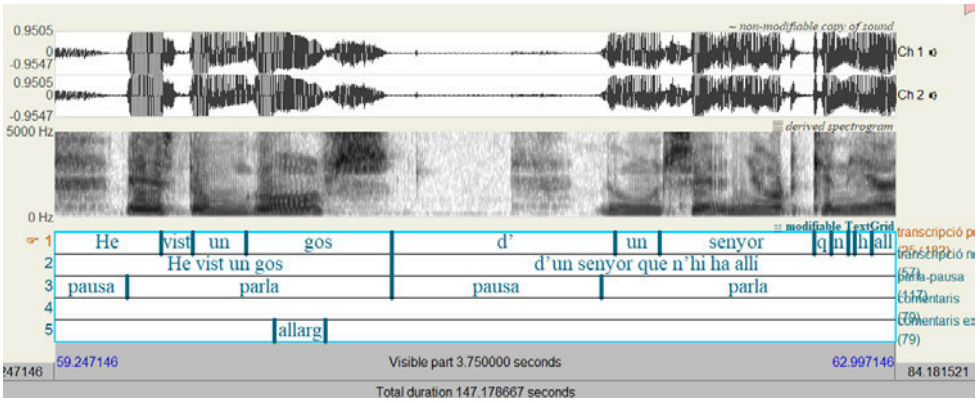
A tall d'exemplificació, tot seguit hi ha una demostració de com transcriu un àudio sense cap ajustament manual en català.

00:03:23.0...	00:03:27.0...	Va, va, les tàcaries al moment del dibuix.
00:03:27.0...	00:03:30.4...	Buこの series, al polarized a countless
00:03:30.4...	00:03:30.4...	on ar
00:03:30.5...	00:03:33.0...	ha una banda que no sé qui et...)
00:03:34.0...	00:03:34.0...	Però volons.
00:03:34.0...	00:03:35.0...	Ja.
00:03:35.0...	00:03:36.0...	Sí, unábbarà.
00:03:37.0...	00:03:38.0...	Ha, ha...
00:03:38.0...	00:03:41.0...	És, m'agom Luoifoo a taxi원, schönes.
00:03:41.0...	00:03:42.0...	Va, è format Hao.
00:03:42.0...	00:03:45.0...	Et la senyor altra, i fiam el mateix.
00:03:45.0...	00:03:47.0...	Què viu, ja?
00:03:48.0...	00:03:54.0...	Aquí ho ia formar t ends a homeless i Initiative de Fourier examples.
00:03:54.0...	00:03:56.0...	Que aquesta 말era,

Es pot observar com sí que hi ha algun fragment detectat adequadament, però, en canvi, molts d'altres que no són correctes. Després, en l'exemple de continuació es pot veure com amb el castellà sí que hi ha una millor detecció de les paraules i de les oracions, tot i que també comet errors.

00:00:00.0...	00:00:18.0...	un árbol que en su cabaña, un coche, los dos que están sentados en el Césper, uno está leyendo el libro
00:00:18.0...	00:00:42.2...	y la otra es baseando una tarnadro, y luego hay un arradio, un acaceror, no sé lo que es, y de este
00:00:42.2...	00:01:07.2...	vamos más adelante, y vemos a uno que está empinando un cachero, el perro está muy atento
00:01:07.2...	00:01:28.6...	y una marca con su pilote, dos vacaciones más allá que estas con vela, que estas llevan
00:01:28.6...	00:01:48.8...	para el de personas, y luego hacemos que uno está, no sé qué recogiendo cargo.

El següent pas de la transcripció l'ha dut a terme una companya del projecte i ha fet ús de l'aplicació PRAAT, que és una eina que permet associar cada paraula amb el so que li correspon, on s'obté la transcripció fonètica associada a l'àudio. Es poden crear diversos nivells d'anàlisi, que en aquest cas, seguint el protocol de transcripció del projecte, han estat cinc: el nivell de transcripció de la paraula; el nivell respiratori, amb les etiquetes d'interlocutor i inintel·ligible; el de la parla i la pausa; el de comentaris que són útils per a l'anàlisi i l'etiqueta de pausa plena i, finalment, l'últim nivell, amb comentaris que poden ajudar a una anàlisi posterior, on es marquen els canvis de codi, les interferències, els titubejos, les autocorreccions, els allargaments, les repeticions i els encavalcaments. En la següent imatge es poden veure els diversos nivells d'anàlisi en un fragment de text d'una transcripció.



Un cop feta la transcripció a l'aplicació de Buzz i després a PRAAT, el següent pas serà enviar les dades al subgrup de computacional, que treballarà per dur a terme les anàlisis estadístiques i les aplicacions informàtiques[5].

[5] Per causes temporals, pel fet que el subgrup de computacional començarà a treballar a posteriori de l'entrega d'aquest treball, s'ha dut a terme una anàlisi computacional pròpia i un marc experimental adaptat a les característiques d'un treball de final de grau de Filologia Catalana.

3.2. *Eines de processament del llenguatge natural: la llibreria de Python spaCy*

L'anàlisi lingüística és un procediment clau per a l'avaluació i tractament de les persones amb patologies neurològiques que afecten la parla. S'estudia la producció en el nivell de l'oració i la paraula, la fluència, les característiques gramaticals, etc. (BRYANT et al., 2016: 504) L'anàlisi es pot fer manualment, però des de l'aparició de la intel·ligència artificial i l'avanç de les aplicacions informàtiques, cada cop són més les eines que permeten automatitzar i agilitzar aquest procés (2016: 505). El camp de la informàtica que s'ocupa de les interaccions entre els ordinadors i el llenguatge humà s'anomena processament del llenguatge natural, o també coneguda amb les sigles PLN. El PLN és definit pel *Termcat* com la disciplina informàtica que permet el tractament computacional de les llengües naturals, per així obtenir, analitzar i generar textos orals o escrits per mitjà d'aplicacions. Els sistemes de PLN són capaços de processar i analitzar elevades quantitats de dades lingüístiques a través de diverses tasques, com la toquenització, l'etiquetatge gramatical o el reconeixement d'entitats anomenades. Un cop processades, se'n poden fer diversos usos, com l'extracció d'informació, l'anàlisi de sentiments (identificar sentiments positius o negatius dins d'un text), el resum automàtic o el reconeixement de veu. Així, el que permeten totes aquestes accions és convertir un text (escrit o oral) amb dades no estructurades, és a dir, un conjunt de lletres i paraules no llegibles pel sistema, a un text amb les dades estructurades, on el contingut està processat i coincideix amb uns patrons reconeixibles.

Els ordinadors treballen amb llenguatges de programació i n'existeixen diversos, cada un dels quals amb funcionalitats i objectius diferents. A través d'aquests, els sistemes són capaços de llegir-ne els codis, interpretar-los i dur a terme les comandes corresponents. En el camp del PLN un dels més usats és Python, que és un llenguatge de programació d'alt nivell amb una multitud d'aplicacions i usos, amb el qual es pot treballar en l'àmbit de l'aprenentatge automàtic, el desenvolupament web o la intel·ligència artificial. Dins dels llenguatges de programació hi ha les llibreries, que són un conjunt de classes i funcions que venen predeterminades en un programari, i que faciliten i ajuden en el procés de codificació als usuaris. Així, no és necessari escriure llargues comandes de codi, perquè la llibreria t'ofereix una varietat de funcions amb línies de codi breus i més senzilles d'emprar. Les llibreries més conegudes de Python dins del terreny del PLN són *Natural Language Toolkit* (o més coneguda com a NLTK), TextBlob o spaCy.

Per poder assolir l'objectiu del treball de determinar l'eficàcia d'una eina de PLN, s'ha plantejat la qüestió de quines llibreries poden ser útils per a l'automatització de l'anàlisi de la parla en català. I també quins ítems són necessaris per a dur a terme una bona anàlisi discursiva d'una persona amb afàsia. Per a resoldre aquesta segona qüestió, s'ha partit dels indicadors discursius de l'obra *Evaluación de la afasia en los bilingües*, de Michel Paradis. A partir d'aquí, s'ha dut a terme una recerca exhaustiva de quines llibreries de PLN obertes, gratuïtes i amb funcionalitats disponibles en català hi ha actualment a Internet. Les llibreries que s'han trobat més destacables són NLTK, Freeling i spaCy. Aquestes tres llibreries són de codi obert i proporcionen recursos per analitzar textos en diverses llengües amb funcionalitats com l'etiquetatge de paraules, la segmentació d'oracions o l'anàlisi morfològica, entre d'altres. Després de buscar informació sobre cada una i fer proves de programació amb textos en català, s'ha optat per treballar amb spaCy. Per una banda, Freeling és una llibreria de C++, que és un altre tipus de llenguatge de programació, i s'ha preferit una llibreria amb el llenguatge Python, com és el cas de NLTK o spaCy. NLTK té algunes funcionalitats en català, com el toquenitzador o el buscador de paraules comunes o *stopwords*. Però s'ha considerat que spaCy és l'opció més accessible i adaptable a les necessitats d'aquest treball, i alhora perquè té més recursos disponibles en aquesta llengua.

spaCy és una llibreria de programari oberta i gratuïta per a fer tasques de PLN a Python. La companyia que la gestiona, Explosion, està especialitzada a desenvolupar eines d'intel·ligència artificial, aprenentatge automàtic i PLN. spaCy permet als usuaris executar una varietat àmplia d'accions per a l'anàlisi computacional de textos en diferents idiomes. Entre les principals funcionalitats, destaquen la desambiguació lèxica (*part-of-speech tagging*), la toquenització, el reconeixement d'entitats anomenades i l'anàlisi de dependències. D'aquesta manera, es poden construir aplicacions que poden processar, extreure informació i entendre grans volums de text[6].

Algunes funcions de spaCy es poden fer de manera independent amb línies pròpies de codi, mentre que d'altres requereixen models entrenats. Els models d'aprenentatge automàtic s'entrenen amb conjunts de dades etiquetades, on aprenen a partir d'exemples que després els serviran per fer prediccions precises i generalitzables a tot allò que processin. Les dades d'entrenament poden ser exemples de text juntament amb el tipus d'etiquetes que es vol que

[6] Vegeu la pàgina web de spaCy: https://spacy.io/usage/spacy-101

predigui el model. Aconseguir un funcionament més robust i eficient dependrà de la quantitat i la qualitat de les dades entrenades. En llengües com l'anglès, les quals disposen d'una quantitat elevada de dades disponibles, és més senzill entrenar aquests sistemes. En canvi, per llengües com el català, amb menys volum de dades disponibles, és més complex aconseguir un rendiment comparable. A més, l'entrenament d'aquests sistemes per a l'anàlisi de discursos de persones amb trastorns de la parla és encara més dificultós, pel fet que és necessari predeterminar de manera acurada quines són les convencions que se segueixen i quines característiques han d'identificar (BRYANT, 2008: 505).

Un element a destacar de l'arquitectura de spaCy és el mòdul anomenat *language*, en el qual s'emmagatzemen regles i característiques pròpies de cada llengua, com per exemple la llista de les paraules més comunes o *stopwords*, les regles de puntuació o les taules de lemes. Dins d'aquest mòdul es troben dos elements imprescindibles perquè es puguin dur a terme les tasques de PLN: el toquenitzador i les *pipelines*. Per una banda, el toquenitzador s'encarrega de dividir el text en toquens, que són les unitats bàsiques amb què es treballa a la lingüística computacional. Poden ser una paraula o un signe de puntuació en un context determinat, inclosos els seus atributs, etiquetes i dependències, i tenen un valor sintàctic i semàntic propi. L'acció del toquenitzador és convertir el text en un objecte *Doc*, que és un text format per toquens iterables que el sistema és capaç de detectar i sobre els quals executarà totes les comandes i operacions. Per a fer-ho, primer divideix el text en caràcters separats per espais en blanc i després aplica les regles incorporades en cada llengua per a comprovar que alguna coincideixi. Per exemple, si es troba una paraula amb guions, depenent de la llengua ho separarà en un o dos toquens. Les *pipelines*, per altra banda, són sèries d'accions dins d'un model preentrenat que el sistema va executant en seqüència (d'aquí el nom de *pipeline*, que vol dir canonada en anglès). Amb cada pas, que és una funció concreta, es modifica el text amb el qual s'està treballant. Aquestes funcions poden ser l'etiquetadora, l'analitzador gramatical, el reconeixement d'entitats conegudes, etc. Els components de les *pipelines* funcionen a partir de models estadístics de spaCy que permeten fer prediccions dins del context, i les seves capacitats de processament depenen dels components, dels models estadístics i de com es va entrenar. A la imatge que es presenta tot seguit es pot observar un fragment de la *pipeline* que s'ha usat en aquest treball, que és la "ca core news lg".

```
{'summary': {'tok2vec': {'assigns': ['doc.tensor'],
   'requires': [],
   'scores': [],
   'retokenizes': False},
  'morphologizer': {'assigns': ['token.morph', 'token.pos'],
   'requires': [],
   'scores': ['pos_acc', 'morph_acc', 'morph_per_feat'],
   'retokenizes': False},
  'parser': {'assigns': ['token.dep',
    'token.head',
    'token.is_sent_start',
    'doc.sents'],
   'requires': [],
   'scores': ['dep_uas',
    'dep_las',
    'dep_las_per_type',
    'sents_p',
    'sents_r',
    'sents_f'],
   'retokenizes': False},
  'attribute_ruler': {'assigns': [],
   'requires': [],
   'scores': [],
   'retokenizes': False},
  'lemmatizer': {'assigns': ['token.lemma'],
   'requires': [],
   'scores': ['lemma_acc'],
   'retokenizes': False},
  'ner': {'assigns': ['doc.ents', 'token.ent_iob', 'token.ent_type'],
   'requires': [],
   'scores': ['ents_f', 'ents_p', 'ents_r', 'ents_per_type'],
   'retokenizes': False}},
```


Per exemple, per poder fer la tasca d'etiquetatge gramatical és necessari descarregar una *pipeline* determinada que tingui disponible el component que permet fer aquesta tasca. S'activarà enviant un determinat codi al conjunt d'instruccions informàtiques (o *script*). Aquest component el que farà és prendre decisions sobre dels toquens de l'objecte *Doc* amb el qual es treballa, a partir de les dades amb les quals el model ha estat entrenat. D'aquesta manera, decidirà quina etiqueta s'ha d'assignar a una paraula determinada.

A la taula següent es recullen les tasques més importants que fa spaCy, amb una breu descripció i el codi a executar.

TASCA	DESCRIPCIÓ
Etiquetatge gramatical o *part-of-speech tagging*	Assigna el tipus de paraules als toquens, sigui un verb, un nom o un adjectiu. **CODI**: toquen.pos_ / toquen.tag

Lematització	Assigna les formes bàsiques de les paraules sense flexions morfològiques o prefixos i sufixos, és a dir, redueix el conjunt de formes flexionades d'un mot a un lema. Per exemple, el lema de "soc" és "ser" i el lema de "rates" és "rata". **CODI**: toquen.lemma_
Segmentació de frases	Troba i segmenta frases individuals. **CODI**: doc.sents
Reconeixement d'entitats anomenades o *name entity recognition*	Etiqueta objectes denominats del món real, com persones, empreses o llocs. **CODI:** doc.ents/ toquen.ent_iob/ toquen.ent_type
Anàlisi morfològica	Assigna les característiques morfològiques dels toquens: cas, gènere, nombre, persona, etc. **CODI:** toquen.morph
Anàlisi de dependència sintàctica	Assigna les etiquetes de dependència sintàctica, que descriuen les relacions entre toquens individuals, per caracteritzar-ne l'estructura gramatical. Aquesta funció ajuda a veure quin rol ocupa cada toquen dins del text. **CODI**: toquen.dep_/ toquen.head/ doc.noun_chunks/ chunk.text/ chunk.root.text/ chunk.root.dep_/ chunk.root.head.text

Un cop entès el funcionament de la llibreria, s'ha dut a terme un procés d'autoaprenentatge a partir de pàgines web i tutorials disponibles a Internet, com el curs de spaCy del doctor William James Mattingly, enginyer de PLN[7]. A més, spaCy té recursos a la seva pàgina web

[7] Vegeu la pàgina web del doctor William James Mattingly: https://spacy.pythonhumanities.com/

que ajuden en aquest procés, com un curs avançat de PLN amb spaCy[8] o guies per a la seva instal·lació o funcionament. Com que totes aquestes pàgines web són principalment en anglès, el que s'ha fet és replicar els guions d'instruccions al català. El pas següent s'ha basat a intentar automatitzar els indicadors de la llista d'ítems d'anàlisi discursiva afàsica de Michel Paradis, a partir de funcions i codis de spaCy[9]. Un cop escrits i executats els diversos codis, i comprovat el seu funcionament amb textos aleatoris, s'han agrupat en els següents cinc apartats, de manera que l'ordre de l'anàlisi tingui un sentit i una coherència.

1. **Informació descriptiva**: en aquest apartat es comptabilitzen el nombre de toquens, d'enunciats, de tipus o *types*, la raó tipus/ toquen i la longitud mitjana dels enunciats.
2. **Fluència lèxica i sintàctica**: aquí es comptabilitzen el nombre de substantius, verbs i preposicions, el nombre de paraules clau, la ràtio de paraules clau entre substantius, les paraules comunes o *stopwords*, el percentatge de preposicions respecte al nombre total de toquens i el nombre de clàusules subordinades.
3. **Disfluència**: terme entès com la dificultat en l'expressió oral continua, que es pot caracteritzar pels següents elements: el nombre total de preguntes sobre el nom d'una paraula, el total d'oracions inacabades o interrupcions del discurs, el total de dubitacions o falques, el percentatge de dubitacions respecte al nombre total de toquens, el total de fragments inintel·ligibles i el percentatge d'aquests respecte al nombre total de toquens, i el total d'ecolàlies.
4. **Ús de les llengües**: s'observa quin ús de les llengües ha fet la persona amb afàsia, característica especialment interessant en contextos bilingües com el del projecte.
5. **Tasques de PLN**: finalment s'executen les principals tasques de PLN que ofereix spaCy, amb la lematització, l'etiquetatge gramatical i morfològic, les dependències sintàctiques, i el reconeixement d'entitats.

3.3. *L'algorisme de spaCy per a l'anàlisi computacional d'un corpus*

A continuació, es presenta l'algorisme, amb els diversos passos a seguir i amb una breu explicació:

[8] Vegeu la pàgina web del curs avançat de PLN que ofereix spaCy de manera gratuïta: https://course.spacy.io/en/
[9] Als annexos hi ha la taula comparativa entre la llista dels indicadors i els codis que es presenten en l'algorisme següent.

S'obre una llibreta a Google Colab[10] i es comença important spaCy:

```
import spacy
```

Després, es descarrega un model en català de spaCy. Els models disponibles es troben a la seva pàgina web:

```
! python -m spacy download ca_core_news_lg
```

Es carrega el model dins la variable "nlp":

```
nlp = spacy.load('ca_core_news_lg')
```

S'observa per quins components està formada la *pipeline* d'aquest model:

```
nlp.analyze_pipes()
```

Es copia el text:

```
text=" "
```

Es converteix el text en un objecte doc, i es processa amb la variable "nlp", on hi ha el model:

```
doc=nlp(text)
```

S'imprimeix el document:

```
print(doc)
```

INFORMACIÓ DESCRIPTIVA

Es calcula el nombre de toquens, de *types* o tipus i la raó entre toquens i tipus. La diferència entre toquen i tipus és que el toquen és cada element individual del text amb una informació pròpia, tal com s'ha descrit anteriorment. En canvi, un *type* es refereix a un toquen únic, és a dir, hi ha la diferència que si un element surt més d'un cop en un text, només es comptabilitza com a un tipus (FundéuRAE). Dit d'una altra manera, els tipus es podrien entendre com el repertori de paraules diferents del text. Per això, s'afegeix la funció *set()* perquè permet eliminar aquells toquens que estan duplicats. Com es pot comprovar a la línia de codi, s'han exclòs els signes de puntuació i les etiquetes *Inintel·ligible*, *Pausa* i *Ecolàlia*, perquè no són toquens pròpiament del discurs de la persona. La raó entre toquen i tipus s'obté de la divisió entre el nombre de paraules diferents dividit entre el nombre de paraules totals. La funció *round* serveix per arrodonir el nombre decimal als dígits que se li demanin.

[10] Vegeu la pàgina web de Google Colab: https://colab.research.google.com/

```
paraules_excloses = ["Inintel·ligible", "Pausa", "Ecolàlia"]
nombre_toquens = len([toquen.text for toquen in doc if not
toquen.is_punct and toquen.text not in paraules_excloses])
nombre_types = len(set([toquen.text for toquen in doc if not
toquen.is_punct and toquen.text not in paraules_excloses]))

print("Nombre de toquens:", nombre_toquens)
print("Nombre de types:", nombre_types)
print("Ràtio type-toquen:", round(nombre_types/nombre_toquens,3),
"("+str(round(nombre_types/nombre_toquens*100,2))+"%)")
```

S'imprimeixen els toquens amb la funció *text.split()*, que permet que els signes de puntuació no es comptabilitzin com a toquens individuals. També s'ha tingut en compte l'exclusió de les paraules *Inintel·ligible*, *Pausa* i *Ecolàlia*:

```
paraules_excloses = ["*Inintel·ligible*", "*Pausa*", "*Ecolàlia*"]
for toquen in text.split():
  if toquen not in paraules_excloses:
    print (toquen)
```

Es calcula el nombre d'enunciats i s'imprimeixen:

```
total_oracions = len(list(doc.sents))
print("Número total d'enunciats:", total_oracions)
for sent in doc.sents:
    print(sent)
```

Es calcula la longitud mitjana dels enunciats. La funció *append* vol dir "adjuntar" i la funció *sort* vol dir "classificar", de manera que s'ordena en ordre ascendent la llista de longitud de les oracions. I es divideix entre dos perquè es calcula l'índex de l'element central en la llista ordenada:

```
sentence_lengths = []
for sent in doc.sents:
    sentence_length = len(sent)
    sentence_lengths.append(sentence_length)
median_sentence_length = sorted(sentence_lengths)[len(sentence_lengths)
// 2]
print("Longitud mitjana dels enunciats:", median_sentence_length)
```

FLUÈNCIA LÈXICA I SINTÀCTICA

S'imprimeix el nombre total de substantius:

```
total_substantius = sum(1 for toquen in doc if toquen.pos_ == "NOUN")
print("Nombre total de substantius:", total_substantius)
substantius=[]
for toquen in doc:
  if toquen.pos_ == "NOUN":
    substantius.append(toquen.text)
print(substantius)
```

S'han determinat paraules clau que es considera que haurien d'aparèixer a la descripció de la imatge, i es comprova quin nombre d'aquestes paraules clau apareixen en la descripció que fa la persona: pícnic, estel, gos, vaixell, barca, casa, arbre, família, bandera, pare, mare, xic i nen. *toquen.text.lower()* és un mètode a Python que converteix una cadena de text a minúscules, d'aquesta manera si alguna d'aquestes paraules comença en majúscula al text, la funció la converteix en minúscula:

```
paraules_clau=["pícnic","estel","gos","vaixell","barca","casa","arbre",
"família","bandera","xic","nen","pare","mare"]
total_paraules = 0
for toquen in doc:
    if toquen.text.lower() in paraules_clau:
        total_paraules += 1
print("Número total de paraules:", total_paraules)
```

S'imprimeixen les paraules clau:

```
paraules_clau_text = set()

for toquen in doc:
    if toquen.text in paraules_clau:
        paraules_clau_text.add(toquen.text)

print("Paraules clau que apareixen al text:")
print(list(paraules_clau_text))
```

S'imprimeix el percentatge de paraules clau respecte al nombre total de substantius. Aquest percentatge equival a la rellevància lèxica dels substantius usats per la persona:

```
resultat_paraules_substantius = (total_paraules/total_substantius)*100
```

```
print("Percentatge de paraules clau respecte al nombre total de substantius:",resultat_paraules_substantius,"%")
```

S'imprimeix el nombre total de *stopwords*, o paraules comunes del català segons spaCy. Les *stopwords* són paraules d'ús freqüent, considerades les primeres que es comencen a produir després d'un episodi agut d'afàsia, i també les que són més fàcilment emeses per persones amb patologies neurològiques que afecten el discurs:

```
num_stopwords = sum(1 for toquen in doc if toquen.is_stop)
print("Número total de stopwords al text:", num_stopwords)
```

```
stopwords_spacy_català = spacy.lang.ca.stop_words.STOP_WORDS
total_stopwords = 0
for toquen in doc:
    if toquen.text.lower() in stopwords_spacy_català:
        total_stopwords += 1
print("Número total de stopwords:", total_stopwords)
```

```
stopwords_spacy_català = spacy.lang.ca.stop_words.STOP_WORDS
stopwords_trobades = []
for toquen in doc:
    if toquen.text.lower() in stopwords_spacy_català:
        stopwords_trobades.append(toquen.text)
print("Stopwords trobades al text:", stopwords_trobades)
```

Es calcula el nombre de verbs. Aquí s'usa la funció suma (*sum()*), que és una altra manera de comptabilitzar el nombre de vegades que apareix un determinat element en un text. Es fa imprimir els verbs per comprovar que el que considera com a verb és adequat:

```
total_verbs = sum(1 for toquen in doc if toquen.pos_ == "VERB")
print("Nombre total de verbs:", total_verbs)
verbs=[]
for toquen in doc:
  if toquen.pos_ == "VERB":
    verbs.append(toquen.text)
print(verbs)
```

S'imprimeix el total de preposicions i quines ha trobat. De la mateixa manera que els verbs, es fa imprimir les preposicions per comprovar que siguin adequades:

```
total_preposicions = sum(1 for toquen in doc if toquen.pos_ == "ADP")
print("Nombre total de preposicions:", total_preposicions)
preposicions=[]
for toquen in doc:
  if toquen.pos_ == "ADP":
    preposicions.append(toquen.text)
print(preposicions)
```

S'imprimeix el percentatge de preposicions respecte al nombre total de toquens:

```
resultat_preposicions = (total_preposicions/nombre_toquens)*100
print("Percentatge de preposicions respecte al nombre total de toquens:",resultat_preposicions,"%")
```

Es calcula el nombre total de clàusules subordinades. Amb la funció *any()* retorna "True" si és veritat la funció que explicita. "advcl" és una etiqueta sintàctica a spaCy que vol dir "clàusula adverbial":

```
comptador_clàusules_subordinades = 0
for sent in doc.sents:
     té_clàusules_subordinades = any(toquen.dep_ == "advcl" for toquen in sent)
    if té_clàusules_subordinades:
        comptador_clàusules_subordinades += 1
print("nombre total de clàusules subordinades:", comptador_clàusules_subordinades)
```

S'imprimeix quines clàusules subordinades ha trobat:

```
for sent in doc.sents:
     té_clàusules_subordinades = any(toquen.dep_ == "advcl" for toquen in sent)
    if té_clàusules_subordinades:
        print(sent.text)
```

DISFLUÈNCIA

Es calcula el total de preguntes que es fa la persona en descriure el text, en el moment que dubta de com es diuen certes paraules. Tanmateix, aquest nombre s'ha de repassar manualment, atès que hi ha preguntes que no són de dubtes sobre certes paraules.

```
total_preguntes = 0
```

```
for toquen in doc:
    if toquen.text == "?":
        total_preguntes += 1
print("Número total de preguntes:", total_preguntes)
```

Es calcula el total d'oracions inacabades o interrupcions del discurs, marcat amb tres punts suspensius en la transcripció:

```
total_trespunts = 0
for toquen in doc:
    if toquen.text == "…":
        total_trespunts += 1
print("Número total de tres punts:", total_trespunts)
```

Total de dubitacions (**Pausa**), per exemple quan emet falques o omplidors, del tipus "ah" o "eh", o amb la presència d'estereotípies verbals:

```
total_pauses = 0
for toquen in doc:
    if toquen.text == "Pausa":
        total_pauses += 1
print("Número total de pauses:", total_pauses)
```

Es calcula el percentatge de pauses respecte al nombre total de toquens. Com que en el moment de dur a terme les gravacions no es va estipular un límit màxim de temps, no es poden comparar els diferents números que han sortit de les quatre persones. Per això, s'han establert proporcions entre nombres de cada pacient:

```
resultat_pauses=(total_pauses/nombre_toquens)*100
print("Percentatge de pauses respecte al total de toquens:",
resultat_pauses,"%")
```

S'imprimeix el nombre de fragments inintel·ligibles, és a dir, moments en què no s'entén el discurs de la persona:

```
total_fragments_inintel·ligibles = 0
for toquen in doc:
    if toquen.text == "Inintel·ligible":
        total_fragments_inintel·ligibles += 1
print("Número total de fragments inintel·ligibles:",
total_fragments_inintel·ligibles)
```

Es calcula el percentatge de fragments inintel·ligibles respecte al nombre total de toquens:

```
resultat_inintel·ligibles=(total_fragments_inintel·ligibles/nombre_toquens)*100
print("Percentatge de fragments inintel·ligibles respecte al total de toquens:", resultat_inintel·ligibles,"%")
```

Es calcula el nombre d'ecolàlies, que marquen el moment que la persona repeteix l'emissió de l'interlocutor:

```
total_ecolàlies = 0
for toquen in doc:
    if toquen.text == "Ecolàlia":
        total_ecolàlies += 1
print("Nombre total d'ecolàlies:", total_ecolàlies)
```

ÚS DE LES LLENGÜES

S'imprimeix l'ús que es fa de les llengües, detectat per la funció *spacy_language_detection*[11]:

```
! pip install spacy-language-detection
from spacy.language import Language
from spacy_language_detection import LanguageDetector

def get_lang_detector(nlp, name):
    return LanguageDetector(seed=42)

Language.factory("language_detector", func=get_lang_detector)
nlp.add_pipe('language_detector', last=True)
doc = nlp(text)

language = doc._.language
print(language)

for i, sent in enumerate(doc.sents):
    print(sent, sent._.language)
```

[11] El codi s'ha extret de la següent pàgina web: https://pypi.org/project/spacy-language-detection/

TASQUES DE PLN

S'imprimeix la lematització dels toquens que ha trobat. Primer es mira que el toquen no sigui el mateix que el lema, amb "!=", de manera que la llista no es faci gaire llarga. Després, s'imprimeix de manera que quedi en una llista. "f" és una manera d'imprimir concreta el text:

```
for toquen in doc:
    if str(toquen) != str(toquen.lemma_):
        print(f"{str(toquen):>20} : {str(toquen.lemma_)}")
```

S'imprimeix la resta de tasques més comunes de PLN que s'ha pogut comprovar que es poden fer amb spaCy i en català: etiquetatge gramatical i morfològic, i dependències sintàctiques. Perquè el resultat no sigui excessivament extens, s'han eliminat les paraules comunes i els signes de puntuació:

```
for toquen in doc:
    if not toquen.is_stop:
        if not toquen.is_punct:
            print(
            f"""
toquen: {str(toquen)}
=====
ETIQUETA: {str(toquen.tag_):10} POS: {toquen.pos_}
EXPLICACIÓ: {spacy.explain(toquen.tag_)}
MORFOLOGIA:{toquen.morph}
DEPENDÈNCIA:{toquen.dep_}"""
    )
```

S'imprimeix un arbre de dependències sintàctiques, amb el *displacy.render*. El *displacy.render* és una funció de spaCy que permet visualitzar els resultats de les tasques de PLN, com les dependències sintàctiques o el reconeixement d'entitats:

```
from spacy import displacy
displacy.render(doc, style="dep")
```

Finalment, s'imprimeix la tasca de reconeixement d'entitats, que també es representa amb el *displacy.render*:

```
for ent in doc.ents:
    print(ent.text,ent.label_, toquen.ent_iob_)
displacy.render(doc,style="ent")
```

VECTORS DE PARAULES

A continuació s'imprimeix un exemple de com funcionen els vectors de paraules del model emprat en aquesta anàlisi, amb la paraula *afàsia*. No forma part pròpiament de l'anàlisi discursiva, però serveix per veure com funciona aquest aspecte a spaCy. És necessari descarregar la llibreria de Python NumPy-, que és una llibreria més adreçada a computació numèrica:

```
import numpy as np
your_word = "afàsia"

ms = nlp.vocab.vectors.most_similar(
         np.asarray([nlp.vocab.vectors[nlp.vocab.strings[your_word]]]),
n=10)
words = [nlp.vocab.strings[w] for w in ms[0][0]]
distances = ms[2]
print(words)
```

Per acabar, es comentaran alguns dels problemes que han sorgit durant la creació de l'algorisme. En primer lloc, ha estat necessari usar la funció *text.split* perquè si no els signes de puntuació es detectaven com a toquens individuals i així s'han pogut incloure com a part d'un toquen. Per segmentar el text en diverses frases s'ha escrit al codi la funció de *list*, per així convertir el *doc.sents* en una llista, ja que per defecte és un "objecte generador" i no es pot comptabilitzar, per això apareix l'error `TypeError: object of type 'generator' has no len()`. La funció *len()* normalment està dissenyada per treballar amb seqüències o col·leccions com les llistes, cadenes de caràcters o *strings*, pel fet que són estructures ben definides i que permeten determinar d'una manera acurada el seu nombre d'elements. Després, les *stopwords* o paraules comunes venen predeterminades per spaCy i en faltarien algunes, tot i que es poden afegir a una *pipeline* pròpia.

Per altra banda, en el moment que s'ha intentat detectar la llengua del text, com que amb spaCy no es pot usar més d'un model al mateix moment, si es fa ús d'un model entrenat en català, detecta tot el text en català, i si és en castellà ho fa en aquesta llengua. Per això ha estat necessari cercar una funció de spaCy que detecti les llengües d'un text, i s'ha trobat la funció *LanguageDetector*, tot i que és una línia de codi molt específica i ha generat força errors a l'hora d'executar-la diverses vegades. Per exemple, el primer cop que s'ha executat s'ha afegit el component a la *pipeline* i després ha estat necessari esborrar la part del codi on

s'afegia la funció. Tanmateix, si s'esborra aquesta part, el sistema perd la referència d'una altra funció, la de *Language*, i llavors ho detecta tot com un error. Una altra qüestió ha estat que a l'hora d'imprimir les paraules clau trobades al text, s'ha hagut d'usar la funció *set()*, perquè d'aquesta manera no es repeteixen les paraules i es troben només els types. La funció *set()* serveix per crear una col·lecció d'elements únics i que no es repeteixen.

Finalment, un cop creat tot l'algorisme, s'han escollit quatre participants bilingües en el projecte amb tres tipus d'afàsia diferent (afàsia de Broca, transcortical motora i anòmica) que han fet l'estudi en català, de manera que així es poden contrastar els resultats obtinguts. S'ha optat per seleccionar dues persones amb afàsia anòmica, les quals tenen graus de gravetats diferents. Hi ha casos de persones amb afàsies i discursos més severs, però s'han descartat, ja que són persones que s'acosten al mutisme i a la producció de vocalitzacions sil·làbiques del tipus "ah!". Després, per aconseguir uns resultats més contrastius i coherents, s'han escollit els fragments de transcripció en els quals les quatre persones descriuen la mateixa imatge, procedent de la *Bateria d'Afàsia Western*, que porta per nom "Escena de pícnic", perquè així es puguin fer evidents les diferències a l'hora de descriure la imatge.

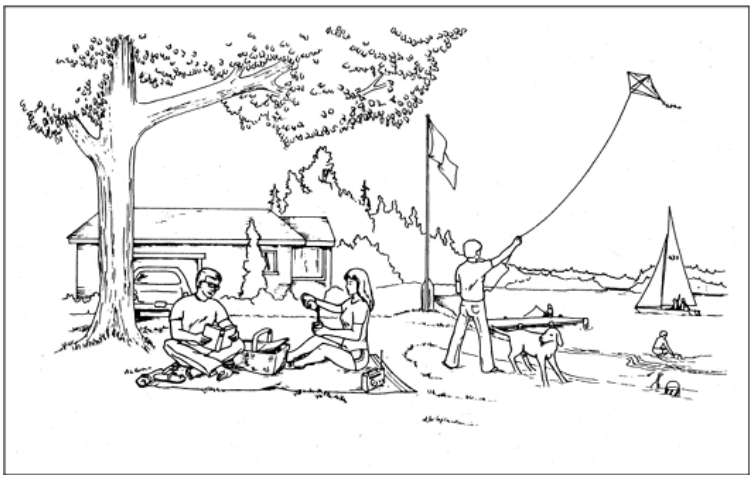


Fig. 2. Dibuix de l'Escena de pícnic. (UGGEN, 2020: 5)

4. RESULTATS

4.1. *Resultats de l'algorisme i el funcionament de spaCy*

Tot seguit es mostraran els resultats que s'han obtingut després d'executar els quatre textos a l'algorisme[12]. Primerament, es comentaran algunes observacions que s'han fet sobre els

[12] Els resultats dels guions d'instruccions es troben en un enllaç a l'*Annex I*, que dirigeix a un repositori de GitHub, un servei d'allotjament de repositoris.

resultats de spaCy, si són adequats o no en català, i quins errors ha comès de manera generalitzada. Després, es presenta una taula amb els resultats quantitatius de l'anàlisi discursiva, amb el nombre de toquens, de substantius, de paraules clau, etc. I, finalment, es mostren algunes observacions sobre els discursos produïts per les persones estudiades, que es fan en funció d'aquests resultats i de les característiques de cada tipus d'afàsia.

La segmentació que fa spaCy dels enunciats en general és adequada, excepte en alguns casos, com per exemple separa una oració[13] en dues a "Suponc que està al costat del mar // perquè n'hi ha una barca, una barqueta". Si s'observen els criteris que segueix en la segmentació quan hi ha tres punts suspensius, en alguns casos ho segmenta com a dues oracions i en d'altres com la mateixa, per tant, no sembla que segueixi un criteri uniforme en aquest cas. Paradis considera que un enunciat pot ser una sola paraula o sintagma, mentre no formi una unitat sintàctica amb el que ve a continuació. Així, en alguns casos els tres punts suspensius sí que han de formar part del mateix enunciat i en d'altres no, tot i que és un criteri difícil d'executar per spaCy, atès que hauria de discernir si és una mateixa unitat sintàctica o no. Després, separa els signes d'interrogació i punts com a enunciats independents. Quan troba paraules entre asteriscos, com *pausa* o *inintel·ligible*, ho deixa com a part de l'enunciat, tot i que hi ha algun cas que l'ha considerat com un enunciat sol.

Pel que fa a les clàusules subordinades, en alguns casos en fa una identificació correcta, però en la majoria són errònies: o es deixa clàusules, com per exemple "n'hi ha un xic que està volant, un, una cometa…". O en d'altres que les considera subordinades quan no ho són, com a "bueno n'hi ha una casa en un cotxe en la porta". Si s'observen els *outputs* del codi que imprimeix les clàusules subordinades, en el text de la persona 2 i 3 detecta adequadament que no n'hi ha. En el de la persona 4 imprimeix correctament dues clàusules, però identifica quatre d'incorrectes. I, finalment, en el discurs de la persona 1 n'identifica tres i només una correcta, de les set presents. Per tant, el resultat de clàusules subordinades s'ha de comprovar amb les oracions que imprimeix perquè el més probable és que sigui erroni. El que també s'ha de repassar manualment són les preguntes que detecta, ja que n'hi ha algunes que són sobre el funcionament de la teràpia i no sobre com es diuen certes paraules, que seria el tipus de pregunta que interessa a l'anàlisi.

[13] S'està usant el terme enunciat com a expressió lingüística mínima d'una idea o d'un fet, seguint la definició del DIEC2, i el terme oració com una forma lingüística amb independència sintàctica i que des d'un punt de vista estructural consta d'un subjecte i d'un predicat.

Algunes de les paraules que considera com a *stopwords* són “hi ha”, “una”, “en”, “un”, “la”, “estan”, “que”, “a”, “perquè”, “no”, “es”, “això” o “el”, entre moltes d’altres. Després, pel que fa a l’ús de les llengües amb la funció *LanguageDetector*, del total de 66 oracions que detecta amb els quatre textos, 46 són detectades amb la llengua correcta i 20 ho fa incorrectament. Per tant, hi ha un 64,44% de les oracions detectades de manera adequada. Els principals errors que fa és en confondre el català i el castellà amb altres llengües romàniques, com l’italià, el francès o el portuguès. Com en els següents casos: “una toalla se dice?” ho detecta com a italià; “com es diu cometa?” ho detecta com a portuguès i “no ho sé, si, si, si…” com a francès.

Finalment, si es para atenció a com ha executat les tasques de PLN, es poden observar diversos resultats. Per una banda, la lematització l’aconsegueix fer de manera bastant adequada en català, excepte en alguns casos com “perquè → porcar”, “veiem→ veiar” o “està→ est”. En la tasca d’etiquetatge gramatical, els substantius els detecta força bé, amb el gènere i el nombre corresponent. Tanmateix, la informació gramatical no és del tot completa en la majoria de les paraules. Reconeix bé el mode de quasi tots els verbs, que és majoritàriament l’indicatiu, i també el nombre, la persona, el temps i la forma, com són els següents casos:

```
toquen: tenen
=====
ETIQUETA: VERB       POS: VERB
EXPLICACIÓ: verb
MORFOLOGIA:Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin
DEPENDÈNCIA:advcl

toquen: volant
=====
ETIQUETA: VERB       POS: VERB
EXPLICACIÓ: verb
MORFOLOGIA:VerbForm=Ger
DEPENDÈNCIA:acl
```

O com aquest pronom:

```
toquen: se
=====
ETIQUETA: PRON       POS: PRON
EXPLICACIÓ: pronoun
MORFOLOGIA:Case=Acc,Dat|Person=3|PrepCase=Npr|PronType=Prs|Reflex=Yes
```

```
DEPENDÈNCIA:obj
```

O com el cas d'aquest participi, que tot i que sigui incorrecte, ja que és una interferència del castellà, el detecta adequadament com a adjectiu i participi:

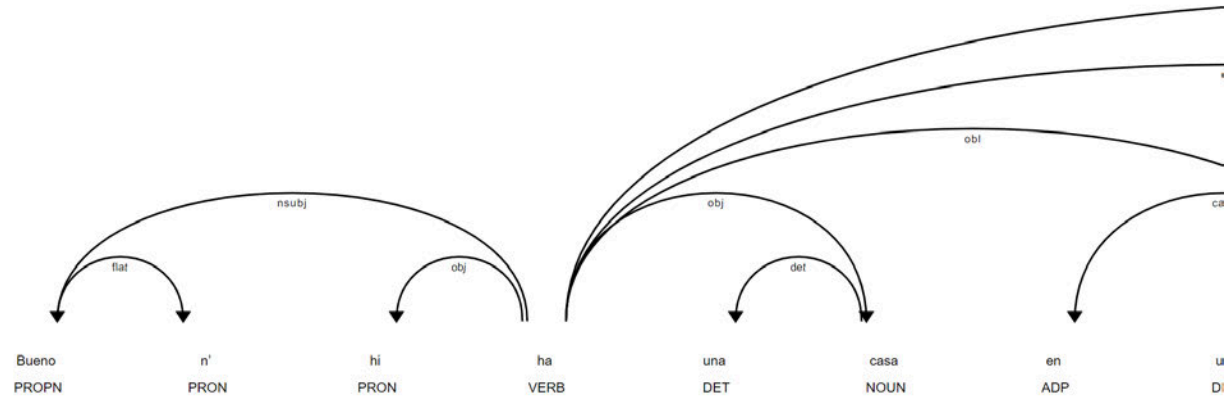
```
toquen: Sentats
=====
ETIQUETA: ADJ        POS: ADJ
EXPLICACIÓ: adjective
MORFOLOGIA:Gender=Masc|Number=Plur|VerbForm=Part
DEPENDÈNCIA:obj
```

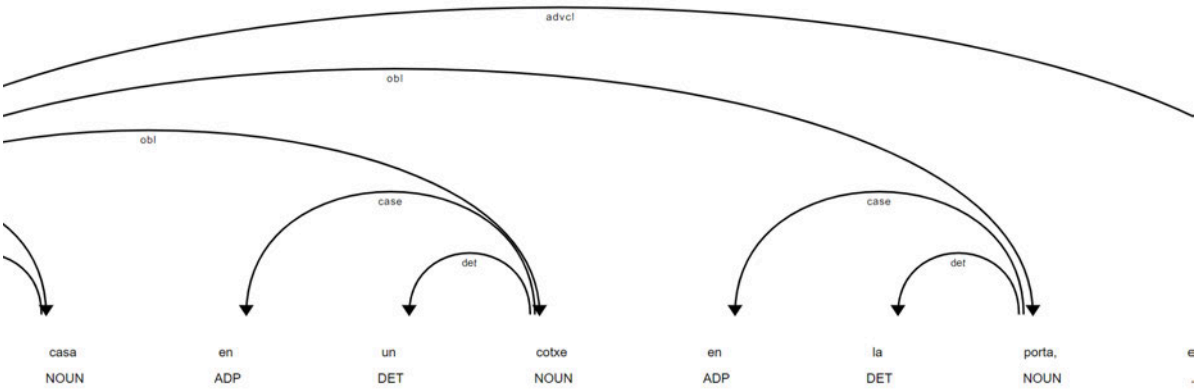
Com a última observació, es podria comentar que sobretot comet errors de detecció en les paraules en castellà, com les següents, que pot ser conseqüència del fet que el model és en català:

```
toquen: dice
=====
ETIQUETA: PROPN      POS: PROPN
EXPLICACIÓ: proper noun
MORFOLOGIA:
DEPENDÈNCIA:conj

toquen: nada
=====
ETIQUETA: PROPN      POS: PROPN
EXPLICACIÓ: proper noun
MORFOLOGIA:
DEPENDÈNCIA:flat
```

En els arbres de dependència sintàctica, tot i ser un discurs amb moltes interrupcions, imprimeix bastant bé les relacions de dependència sintàctica, malgrat que els errors són superiors a les relacions detectades correctament. Per tant, en el cas del català és una funció a millorar notablement. A continuació, es mostra un fragment (dividit en dos per l'espai de la pàgina) d'un arbre de dependència d'un text, amb la funció *displacy.render*:





En aquest fragment en concret es pot constatar com reconeix adequadament la relació de dependència en dos sintagmes preposicionals, atès que "un cotxe" i "la porta" depenen de la preposició "en" respectivament. Per altra banda, imprimeix de manera correcta les relacions de determinació dels substantius amb els determinants o els complements directes. Tanmateix, comet alguns errors en l'atribució de la funció de subjecte del pronom relatiu "que" a una oració subordinada, com en el cas de "suponc que està al costat", quan és funció de complement directe. Però sí que detecta bé la funció de subjecte en el cas de "hi ha un xic que està volant un estel". O també és capaç de separar a + el a "al", i de + el a "del".

En darrer terme, l'última funció a comentar és la de reconeixement d'entitats anomenades. En català és una tasca que no funciona, atès que no reconeix cap etiqueta d'entitat, del tipus *ORG* (organització) o *GEO* (localització), com sí que fa en llengües com l'anglès o el castellà. Sí que assigna a algunes paraules l'entitat de *MISC*, que equival a miscel·lània, és a dir, un conjunt de diferents tipus d'entitats. Però els assigna a toquens com "una", "suponc", "eso", o

"aquí". Altres casos erronis són els de col·locar l'etiqueta *ORG* a "estel" o l'etiqueta *PER* (*persona*) a "bueno". A més, no reconeix entitats ni tampoc toquens amb múltiples components. A tall de recapitulació, de les tasques de PLN, el reconeixement d'entitats és la tasca on comet més equivocacions. A l'arbre de dependències sintàctiques també en fa, tot i que és capaç de detectar adequadament algunes relacions sintàctiques. I pel que fa a l'etiquetatge gramatical i morfològic, en general prou bé, però també comet alguns errors, especialment en els verbs.

4.2. *Resultats de l'anàlisi discursiva del corpus*

A la taula següent es presenten els resultats numèrics de l'anàlisi discursiva i observacions de cada persona analitzada en funció dels seus resultats i de les característiques del tipus d'afàsia.

ÍTEMS	PERSONA 1	PERSONA 2	PERSONA 3	PERSONA 4
INFORMACIÓ DESCRIPTIVA				
Nombre de toquens	194	60	27	128
Nombre de tipus	101	41	21	79
Raó tipus-toquen	52%	68,33%	77,78%	61,72%
Nombre d'enunciats	26	18	4	20
Longitud mitjana dels enunciats	9	5	16	6
FLUÏDESA LÈXICA I SINTÀCTICA				
Nombre de substantius	30	9	6	18
Nombre de paraules clau que apareixen a la descripció	10	4	0	6
Percentatge de paraules clau	33,33%	44,44%	0%	33,33%

respecte al nombre total de substantius				
Nombre de *stopwords* o paraules comunes del català	97	27	14	77
Nombre de verbs	19	10	4	19
Nombre de preposicions	8	5	4	9
Percentatge de preposicions respecte al nombre total de toquens	4%	7,57%	11,11%	6,97%
Nombre de clàusules subordinades	3	0	0	6
DISFLUÈNCIA				
Total de preguntes	5	2	0	2
Total d'oracions inacabades o interrupcions del discurs	22	10	2	17
Total de dubitacions, falques o estereotípies verbals (*Pausa*)	6	5	5	0
Percentatge de pauses respecte al total de toquens	3%	7,57%	13,88%	0%
Nombre de fragments inintel·ligibles (*Inintel·ligible*)	0	0	4	0
Percentatge de fragments inintel·ligibles respecte al total de toquens	0%	0%	11,11%	0%
Nombre total d'ecolàlies (*Ecolàlia*)	0	1	0	1

PERSONA 1

Tipus d'afàsia: **afàsia transcortical motora**[14]**,** d'un ictus isquèmic del novembre del 2018.

Text transcrit:

*Bueno n'hi ha una casa en un ca, cotxe en la porta, estan, tenen un... ai, una toalla... Una... Una toalla se dice? Sí... *Pausa* Tenen un pícnic, una ràdio... *Pausa* Suponc que està al costat del mar perquè n'hi ha una barca, una barqueta. N'hi ha un xic que està vol... *Pausa* està volant un, una cometa... una... una... ai, como se dice en valenciano, la... bandera, como se dice en valenciano? Un pendó, porque es un pendó, no? Lo llaman eso. He vist un gos d'un senyor que n'hi ha allí, suponc que està agarrant algo. *Pausa* Un arbre pareix que estiga un...*Pausa* Això és un, eso es una... Bueno pues nada más. Otro, un altre arbre ahí, el xic duu gafes, ell està ficant algo en un ba... como se dice... No lo se... La test, la... Esta, no. Esta es? *Pausa* No... El pícnic que suponc que s'ho ha menjat. *Pausa* Pues no... No veig res, que... que allí n'hi ha una persona allí també, pareix que sea un...*Pausa* Això pot ser, o no... Me voy a poner las gafas... N'hi ha algo més?*

Aquesta persona fa ús de deu paraules clau i el percentatge d'aquestes respecte al nombre de substantius (trenta) és d'un 33,33%, valor que es pot considerar adequat. L'ús de verbs és relativament divers, i sobretot empra l'indicatiu i el gerundi, que són els temps que normalment es fan servir per a les descripcions d'imatges. Dels 194 toquens, 97 són paraules comunes o *stopwords*, per tant, quasi un 50% de les paraules són considerades com a no comunes en el català. Presenta dificultats en la denominació i això provoca estructures verbals incompletes i latències per buscar les paraules desitjades. Hi ha un total de vint oracions inacabades i es fa tres preguntes de com es diuen certes coses durant el discurs. Fa sis dubitacions, que sobretot es troben al final[15]. És un indici de com se li fa cada cop més complicada l'execució del discurs i com té més dificultats per descriure la imatge, per la incapacitat de trobar les paraules apropiades per fer-ho. Així, al final, són més evidents les conductes d'aproximació, on la persona fa múltiples intents en els quals es va autocorregint la producció per intentar oferir la resposta correcta (DIÉGUEZ i PEÑA, 2012: 289). Diéguez

[14] A la classificació que fa Faustino al seu llibre distingeix entre afàsia transcortical motora clàssica i dinàmica. Com que al diagnòstic els logopedes no fan distinció entre aquests dos tipus, aquí es considera com a afàsia motora transcortical.

[15] Això és una observació que s'ha de fer manualment, atès que spaCy no et detecta exactament a on es troben.

remarca de l'afàsia motora transcortical el fet que l'individu, davant de l'acte de descriure una làmina, no aconsegueix un llenguatge desenvolupat i coherent (2012: 106), tal com es com comprovar en la descripció d'aquesta persona. Hi ha una elevada presència de frases inacabades i d'anòmia, que és la incapacitat d'accés al lèxic i que causa una sensació de llenguatge agramàtic, especialment al final del discurs (AGUADO et al., 2013: 43).

Tanmateix, tot i que l'expressió es realitza amb esforç, és una parla relativament fluent amb alguna oració completa, ben articulada i amb absència de parafàsies. En aquest cas es denota clarament la bona evolució de la teràpia des de la data de l'ictus (fa uns cinc anys aproximadament). És un tipus d'afàsia que si es millora, pot evolucionar a una afàsia anòmica, que podria ser el cas d'aquesta persona. En darrer terme, pel que fa a les interferències lingüístiques, és una persona que la seva llengua habitual és el castellà i aquest fet es fa evident quan dubta dels noms en català, i també perquè li surten força expressions en castellà.

PERSONA 2
Tipus d'afàsia: **afàsia anòmica,** d'un ictus hemorràgic del gener del 2023.
Text transcrit: *Veiem… *Pausa* Jugar a la cometa, com es diu? Com es diu cometa? Eso, *Ecolàlia* L'estel. L'estel blau. L'estel… *Pausa* Estan… Estan llegint. L'home, la dona, estan a… Sentats a… A una manta… De càmping. No ho sé, sí, sí… *Pausa* Un lago, un llac… *Pausa* No ho sé. *Pausa* La platja, estan jugant. Sí, el gos… I… Bueno, no ho sé.*

Dels nou substantius que usa, quatre són paraules clau, que són "gos" i "estel". Dels 60 toquens, 27 són considerats *stopwords*. Fa un ús limitat de preposicions, ja que només diu "a" i un cop "de", i un ús reduït de verbs, pel fet que empra bàsicament "diu", "sé", "veiem" i alguns gerundis, com "llegint" o "jugant". És un discurs poc fluid, buit i interromput, amb presència d'agramatisme, i amb aproximacions a una parla telegràfica. Hi ha frases inacabades i latències en la selecció de les paraules, pel fet que presenta una evident limitació d'accés al lèxic. L'anòmia és la característica més destacable en l'afàsia anòmica, per la dificultat de trobar les paraules durant el discurs espontani. Es fa palès que el castellà és la llengua d'ús habitual de la persona, per la presència d'interferències i perquè en un primer

moment li surt el nom en aquesta llengua, i dubta en català. Fa alguns calcs del castellà, com per exemple amb l'ús de "sentats". Es podria considerar un cas d'arxenofasia, en el qual es fa ús d'una altra llengua, la seva primera en aquest cas, per denominar un element de la imatge (JUNQUÉ, 1994: 93). No hi ha presència de cap circumloqui, però sí d'una ecolàlia. Tanmateix, és un discurs ben articulat, pel fet que no es detecta cap fragment inintel·ligible.

PERSONA 3
Tipus d'afàsia: **afàsia anòmica,** de causa desconeguda de l'abril del 2021.
Text transcrit: *Vale, hi ha una parella al veïnat... Jo no ho veig, entre cometes *Inintel·ligible* *Pausa* La parella *Inintel·ligible* terra *Pausa* i... *Pausa* Que ballen al bosc. *Pausa* *Inintel·ligible* Estan al terra *Inintel·ligible* *Pausa**

Aquest discurs no té quasi enunciats i no s'ha detectat l'ús de cap paraula clau. No obstant això, convé tenir en compte que la persona empra alguns sinònims adequats per a la descripció, com "parella", "veïnat" o "bosc", i es podria considerar una limitació del codi, atès que actualment no és possible trobar sinònims en català amb spaCy. La resta de discurs és inintel·ligible (s'han detectat quatre fragments inintel·ligibles). És el discurs més afectat i incoherent, no només per la inexistència d'oracions, sinó també per la manca de sentit d'algunes clàusules. L'etiqueta diagnòstica de la persona és una afàsia anòmica, però és un discurs amb problemes greus de coherència i de producció, i no fluent. Per tant, es podria afirmar que no s'adequa del tot a les característiques del tipus d'afàsia diagnosticada. Tampoc es troba en una fase aguda, ja que fa dos anys de l'aparició de la patologia (el diagnòstic diu que és de causa desconeguda). Pel que fa a l'ús de les llengües, només es detecta l'ús de català.

PERSONA 4
Tipus d'afàsia: **afàsia motora o de Broca,** per ictus hemorràgic el desembre del 2022.
Text transcrit:

*Aquí pues es veu un dia de camp. I que està... Pos, pos jugant al camp, la família, un fa una cosa, un altre una altra... Què fan? Pues mira el pare es veu aquí llegint el llibre, la mare està fent algo com de berenar, el fill d'aquest està jugant amb una... Ara no m'enrecordo com es diu... *Ecolàlia* Sí, i coses així. Es veu el... vaixell. I aquí pues... Una... Estan amb una... Com es diu? Bueno, no sé, amb un... amb una... D'això... Amb una... Ai, no me'n surto. Bueno... Amb una... No me'n surto. Bueno si et podria destacar aquí hi ha una... Una bandera, que no sé qui és però bueno. I un arbre... I... Quatre coses més.*

Fa ús tant de substantius com de verbs, tot i que presenta una evident dificultat d'accés al lèxic (anòmia), que provoca un discurs poc fluid, simple i molt interromput. La majoria d'oracions són breus, inacabades i incompletes, però articula bé els mots. Hi ha presència d'alguna clàusula subordinada i cap al final esdevé una parla més telegràfica, amb esforç per produir-la. Tot i estar en una fase bastant aguda de la patologia, encara és capaç de produir algunes oracions amb subjecte i predicat, del tipus "pues mira el pare es veu aquí llegint el llibre". No hi ha una supressió completa del llenguatge, ni vocalitzacions inarticulades, ni estereotípies del tipus consonant-vocal o consonant-vocal-consonant (DIÉGUEZ i PEÑA, 2012: 109) o mutisme (ARDILLA, 2013: 9). Fa ús de set paraules clau en la descripció, i de paraules òmnibus del tipus "altre" i "cosa" quan no li surten les paraules, o expressions com "no sé", "no me'n surto", o preguntes com "com es diu?" o "què fan?". Dels 128 toquens, 77 són *stopwords*. Pel que fa a l'ús de les llengües, a part d'algun petit exemple com amb el "pues", es demostra que la persona té el català com a llengua dominant, ja que en comparació amb els altres discursos, aquí no hi ha quasi interferències del castellà.

5. CONCLUSIONS

L'afàsia és un trastorn del llenguatge que afecta de manera evident a la qualitat de vida de les persones que la pateixen, atès que els impedeix desenvolupar-se de manera efectiva en l'acte comunicatiu. La parla no només serveix per a l'intercanvi d'informació, sinó també permet a les persones relacionar-se i expressar les seves emocions, inquietuds i sentiments. És per això que esdevé imprescindible el procés de rehabilitació i tractament de les persones que pateixen afàsia o altres patologies que afecten en la parla. Al llarg del treball s'ha posat de manifest la importància de l'anàlisi de les capacitats lingüístiques de la persona com a un dels indicadors

clau per a l'estudi de la lesió cerebral. Conèixer les característiques que descriuen els diferents tipus d'afàsia que existeixen poden ajudar a oferir una teràpia més individualitzada a les necessitats de cada cas, i així obtenir uns millors resultats.

Altrament, esdevé essencial fer ús d'avaluacions del llenguatge amb instruments de mesura objectius i adaptats a les característiques específiques de cada llengua. I sobretot en contextos com els d'aquest treball, en el qual hi ha la condició de bilingüisme i on l'estudi sistemàtic de les llengües conegudes per l'individu és un requisit per dur a terme una bona rehabilitació. Per tot això, la tasca de Michel Paradis amb el seu *Test d'afàsia per a bilingües* ha estat una tasca destacada per a l'estudi de l'afàsia en aquest context. A partir de la bibliografia emprada pel treball, s'ha pogut constatar que actualment hi ha una quantitat notable d'estudis sobre les afàsies i sobre la seva classificació. També s'han pogut conèixer els exàmens més destacats de l'avaluació de l'afàsia, com el *Test de Barcelona*, el test de Michel Paradis o la *Bateria d'Afàsies Western*, que és d'on s'ha extret la imatge amb la qual s'ha fet l'estudi als participants del projecte de la UPC i l'IEC. Aquestes avaluacions es poden fer manualment, però gràcies a l'evolució de les eines de PLN, s'ha pogut comprovar durant el treball que és possible automatitzar alguns dels seus processos i tasques, i així dur a terme una anàlisi més breu i eficaç.

La realització d'aquest treball ha permès a l'autora participar en el desenvolupament d'un projecte de recerca universitari amb dades clíniques, i ser coneixedora de les complicacions i reptes inherents en aquest tipus de projectes. Una de les principals complexitats ha estat la llengua objecte d'estudi, el català, atès que la quantitat de dades és més limitada que en altres llengües. Això s'evidencia amb el fet que s'han transcrit quasi tres hores més d'àudio en castellà que en català. Per altra banda, obtenir participants que puguin fer l'estudi en català ha estat i continua sent un dels reptes del projecte, especialment en contextos amb una regressió evident d'aquesta llengua, com succeeix a la ciutat de València i de l'Hospitalet de Llobregat. A més, convé afegir que es treballa amb parlants afàsics, que presenten unes característiques més complexes que una parla no patològica. No obstant això, el procés de recerca i d'assoliment dels objectius del projecte es manté actiu, i s'han proposat altres línies d'investigació, com per exemple l'estudi de pauses en discursos quasi inintel·ligibles. Per altra banda, l'autora del treball ha participat en les reunions de treball del projecte i també

apareix com a autora en l'elaboració d'un pòster per al congrés *Interdisciplinary Advances in Statistical Learning*, celebrat del 5 al 7 de juny de 2024, a Donostia (País Basc)[16].

S'han transcrit aproximadament tres hores i mitja d'àudio dels participants en el projecte. Tanmateix, el programa emprat per a l'automatització de la transcripció ha comportat força problemàtiques en les transcripcions de català, atès que ha estat necessari refer quasi tots els textos de nou. Això es deu al fet que no ha pogut detectar adequadament aquesta llengua en moltes ocasions, i l'ha confós amb d'altres. D'aquesta manera, s'ha fet evident la manca d'entrenament i de dades en català del programa, i fa palesa la limitació pròpia de les eines de reconeixement de veu basades en intel·ligència artificial en aquesta llengua. En canvi, el procés de transcripció en castellà ha estat molt més ràpid perquè l'ha detectat de manera òptima. Igualment, la verificació manual de les transcripcions esdevé imprescindible perquè en aquest tipus d'estudis sempre se segueix un protocol de transcripció, amb unes convencions per a situacions específiques, com la transcripció de les pauses o dels errors gramaticals. Un protocol ben definit permet una major eficiència i rapidesa en l'anàlisi discursiva.

S'ha dut a terme una cerca extensa de llibreries amb les quals és possible analitzar textos en català per poder assolir l'objectiu de determinar l'eficàcia d'una eina de PLN per analitzar un corpus d'individus afàsics en aquesta llengua. Entre les opcions més destacades s'han trobat les llibreries de Python NLTK i spaCy, i Freeling, que treballa amb el llenguatge de programació C++. Després de fer proves de programació amb totes tres i valorar les necessitats d'aquest treball, s'ha acabat optant per dur a terme l'algorisme amb spaCy. Una de les raons ha estat que és una llibreria que disposa de nombrosos tutorials i pàgines web que permeten l'aprenentatge autònom del seu funcionament i de l'elaboració dels seus codis. S'ha entès en profunditat la seva arquitectura, amb els diversos mòduls que formen spaCy: els models preentrenats, el mòdul *language*, les *pipelines* i l'objecte *Doc*, entre d'altres. Tanmateix, una dificultat que convé destacar és que totes aquestes pàgines web són en anglès i, per això, ha estat necessari replicar i adaptar els guions d'instruccions al català.

Per a crear l'algorisme s'ha partit dels indicadors de l'anàlisi discursiva del *Test d'afàsia per a bilingües* de Michel Paradis i s'ha intentat automatitzar el màxim nombre d'aquests amb

[16] Publicat a la següent pàgina web: https://upcommons.upc.edu/handle/2117/409243

línies de codi de spaCy. En general, s'ha aconseguit en un bon conjunt d'ítems, especialment per comptabilitzar nombre de verbs, substantius, paraules clau o oracions, i imprimir-los. El guió d'instruccions s'ha dividit en cinc apartats amb les diferents línies de codi per cada indicador: primerament, la informació descriptiva, amb el nombre de toquens, d'enunciats o la raó tipus/toquen. Després, els codis que determinen la fluència lèxica i sintàctica, com el nombre de substantius, verbs o paraules clau, i els que, per contrari, caracteritzen la disfluència, on es pot trobar el total d'oracions inacabades o de dubitacions. I, en darrer lloc, l'ús de les llengües que ha fet l'individu i les tasques de PLN (lematització, etiquetatge gramatical i morfològic, dependències sintàctiques, i reconeixement d'entitats). Tanmateix, hi ha altres indicadors que no ha estat possible automatitzar, com els relacionats amb la coherència del discurs, ja que és complex poder-ho recollir de manera objectiva a través d'un codi, o les omissions de morfemes gramaticals o de pronoms. Tampoc els errors que comet la persona, sigui amb la producció de paraules errònies o en l'ordre de les paraules. Finalment, no s'han tingut en compte qüestions prosòdiques, atès que el que s'ha processat són textos escrits, ni s'ha fet una valoració del tipus de paraules i verbs que empra, pel fet que és una tasca que s'hauria de fer manualment.

D'altra banda, les limitacions del funcionament de spaCy en català són evidents respecte a altres llengües com l'anglès o el castellà, principalment per la manca d'entrenament i dades dels models d'aprenentatge automàtic en aquesta llengua. Nombroses funcions que té spaCy no estan disponibles en català, com la de cercar sinònims, que hagués estat útil per calcular el nombre real de paraules clau. I s'han produït força errors, com per exemple a l'hora de detectar les clàusules subordinades o en segmentar alguns enunciats. En les comandes on s'han detectat més errors són en les eines de PLN de spaCy, especialment en el reconeixement d'entitats anomenades i l'arbre de dependències sintàctiques. La funció de detecció de llengües ha obtingut uns resultats prou adequats, tot i les confusions amb altres llengües romàniques. Per altra banda, l'anàlisi discursiva que s'ha fet de cada persona s'ha adequat bastant a les característiques dels tres tipus d'afàsia (Broca, transcortical motora i anòmica). Excepte un dels casos d'afàsia anòmica, que s'ha considerat que potser s'allunya bastant dels trets descrits en la bibliografia.

Per concloure, s'ha pogut evidenciar que les eines de PLN permeten agilitzar certes tasques del procés d'avaluació lingüística de la parla afàsica, com per exemple a l'hora de comptabilitzar verbs o substantius, o en buscar patrons concrets. Tanmateix, els recursos que

hi ha disponibles actualment en català presenten certes limitacions per executar anàlisis completes, a causa del reduït nombre de recursos disponibles i per la manca de dades anotades. Així doncs, projectes com el de la UPC i l'IEC en el qual s'ha participat, o també el del projecte Aina[17], impulsat per la Generalitat de Catalunya, poden ajudar a disposar de recursos computacionals més eficients i productius en aquesta llengua. Per així aconseguir desenvolupar un ventall més ampli d'aplicacions informàtiques que millorin les teràpies i els tractaments per a persones amb afàsia i altres patologies del llenguatge, com l'Alzheimer o el Parkinson. I conseqüentment poder assolir una recuperació més ràpida de la capacitat comunicativa i funcional, i una millora en la seva qualitat de vida.

[17] Vegeu pàgina web del Projecte Aina i spaCy: https://aina.bsc.es/apps/spacy

6. REFERÈNCIES BIBLIOGRÀFIQUES

AGUADO, G., CARDONA, M., SANZ-TORRENT M., i ANDREU, L. (2013). *Trastorns del llenguatge oral*. Repositori de la Universitat Oberta de Catalunya. Disponible a: https://openaccess.uoc.edu/handle/10609/70965 [última consulta: 11 de maig de 2024]

ARDILA, A. (2013). *Trastornos del lenguaje en pacientes con lesiones cerebrales*. Florida International University: Miami, Florida, USA. Disponible a: https://aalfredoardila.wordpress.com/wp-content/uploads/2013/07/2013-ardila-trastornos_del_lenguaje_en_pacientes_con_lesiones-cerebrales.pdf [última consulta: 25 d'abril de 2024]

BALL, M.J. (2008). *The Handbook of clinical linguistics*. Malden: Blackwell.

BRYANT L., FERGUSON A., i SPENCER E. (2016). "Linguistic analysis of discourse in aphasia: A review of the literature". *Clin Linguist Phon*, 30, pàgines 489-518. Disponible a: https://pubmed.ncbi.nlm.nih.gov/27002416/ [última consulta: 10 de maig de 2024]

BERMÚDEZ, M. (2023). *Las esterotípias verbales en personas con daño cerebral adquirido*. Hermanas hospitalarias. Red Menni de Daño cerebral. Disponible a: https://xn--daocerebral-2db.es/ca/publicacion/las-estereotipias-verbales-en-personas-con-dano-cerebral-adquirido/ [última consulta: 11 d'abril de 2024].

COLL-FLORIT, M. (2013). *Llenguatge, ment i cervell*. Barcelona: Editorial UOC.

DIÉGUEZ-VIDE, F., i PEÑA-CASANOVA, J. (2012). *Cerebro y lenguaje: sintomatología neurolingüística* (1a ed.). Madrid: Editorial Médica Panamericana.

FERNÁNDEZ, Y. (2023). *OpenAI Whisper: qué es, cómo funciona y cómo puedes usar esta inteligencia artificial para transcribir audios*. Xataka Basics. Disponible a: https://www.xataka.com/basics/openai-whisper-que-como-funciona-como-puedes-usar-esta-inteligencia-artificial-para-transcribir-audi [última consulta: 23 d'abril de 2024]

FreeCodeCamp.org. (2022) Natural Language Processing with spaCy & Python - Course for Beginners. Disponible a: https://www.youtube.com/watch?v=dIUTsFT2MeQ [última consulta: 12 d'abril de 2024]

FundéuRAE. Consideraciones teóricas en torno a la riqueza lingüística. Disponible a: https://www.fundeu.es/consideraciones-teoricas/ [última consulta: 14 d'abril de 2024]

GeneracionIA. *Buzz IA. Guia completa. Transcripción y traducción de audio*. Disponible a: https://generacionia.com/audiosia/buzz-ia-guia-completa-transcripcion-y-traduccion-de-audio/ [última consulta: 23 d'abril de 2024]

GÓMEZ, M.I. (2013). *Aplicabilidad del test de la afasia para bilingües de Michel Paradis a la población catalano/castellano parlante*. Barcelona: Universitat de Barcelona. Disponible a: https://diposit.ub.edu/dspace/handle/2445/42564 [última consulta: 23 de maig de 2024]

Institut d'Estudis Catalans. *Diccionari de l'Institut d'Estudis Catalans*. Disponible a: https://dlc.iec.cat/

Institut d'Estudis Catalans. Secció de ciències i tecnologia. Semàntica de les paraules del català: teoria i aplicacions clíniques. Disponible a: https://iec.cat/recerca-v/projecte1.asp?codi=PRO2022-S03-HERNANDEZ [última consulta: 9 de maig de 2024]

Institut d'Estudis Catalans. Secretaria científica. (2023) *Programes de recerca desenvolupats durant l'any 2023. Informe anual. Semàntica de les paraules del català: teoria i aplicacions clíniques*.

JUNQUÉ, C., i BARROSO, J. (1994). *Neuropsicología*. Madrid: Síntesis.

MATTINGLY, W.J.B. (2021) *Introduction to spaCy*. Smithsonian Data Science Lab and United States Holocaust Memorial Museum. Disponible a: https://spacy.pythonhumanities.com/intro.html [última consulta: 2 de maig de 2024]

MIR, M. (2022). *La cara oblidada de les paraules: avaluació funcional del llenguatge en afàsia*. Universitat Oberta de Catalunya, Universitat de Vic, Universitat Central de Catalunya. Disponible a: https://openaccess.uoc.edu/handle/10609/146769?locale=es [última consulta: 13 de maig de 2024]

PARADIS, M., i LIBBEN, G. (1993). *Evaluación de la afasia en los bilingües*. Barcelona: Masson.

Pypi.org. *Spacy language detection 0.2.1*. Disponible a: https://pypi.org/project/spacy-language-detection/ [última consulta: 18 d'abril de 2024]

Softcatalà. *Buzz. Transcripció amb Whisper, d'OpenAI*. Disponible a: https://www.softcatala.org/programes/buzz/ [última consulta: 23 d'abril de 2024]

spaCy. spaCy 101: everything you need to know. Disponible a: https://spacy.io/usage/spacy-101 [última consulta: 9 de maig de 2024]

TORRE, I.G., ROMERO, M., i ÁLVAREZ, A. (2021) "Improving aphasic speech recognition by using novel semi-supervides learning methods on AphasiaBank for English and Spanish". *Applied Sciences*, 11. Disponible a: https://doi.org/10.3390/app11198872 [última consulta: 12 de maig de 2024]

UGGEN, T. K. E. (2020). *The Use of Machine Learning Algorithms and Statistical Models to Classify Aphasia Severity*. University of Technology Sydney. School of Mathematical and Physical Sciences. Disponible a: https://opus.lib.uts.edu.au/bitstream/10453/143921/2/02whole.pdf [última consulta: 29 d'abril de 2024]

Universitat Politècnica de Catalunya. Futur UPC. *Semàntica de les paraules del català: teoria i aplicacions clíniques*. Disponible a: https://futur.upc.edu/32847767 [última consulta: 13 de juny de 2024]

Universitat Oberta de Catalunya. Espai de recursos de dades de ciència de dades. *Processament del llenguatge natural*. Disponible a:

https://datascience.recursos.uoc.edu/processament-del-llenguatge-natural-nlp/ [última consulta: 13 de maig de 2024]

VEGAS, E. (2010) *Preprocessament de les dades*. Universitat Oberta de Catalunya. Disponible a: https://openaccess.uoc.edu/bitstream/10609/141547/7/Preprocessament%20de%20les%20dades.pdf

ZARAGOZA M. et al. (2022). Adaptació de la Bateria d'Afàsies Western al català. Quadern d'estímuls en català. Disponible a: https://futur.upc.edu/34265370 [última consulta: 2 de juny de 2024]

WEBB, W.G., i ADLER, R.K. (2010). *Neurología para el logopeda* (5a ed.) Barcelona: Elsevier.

WRIGHT, H. (2011). *Discourse in aphasia: An introduction to current research and future directions*. Aphasiology, 25, pàgines 1283-1285. Disponible a: *https://www.tandfonline.com/doi/full/10.1080/02687038.2011.613452* [última consulta: 28 d'abril de 2024]

7. ANNEXOS

7.1. *Annex I:*

Enllaç a la pàgina web del **GitHub** de l'autora amb els quatre *scripts* o guions d'instruccions, amb els noms Persona 1, Persona 2, Persona 3 i Persona 4, i amb l'*abstract* del treball final de grau: https://github.com/nuriapoch/TFG-Nuria-.git

7.2. *Annex II:*

GLOSSARI DE TERMINOLOGIA DE L'AFÀSIA

1. **Estereotípia verbal**: són sons, síl·labes o conjunt de paraules que la persona emet repetidament quan intenta articular el llenguatge en qualsevol situació comunicativa. Pot evolucionar com un quadre de mutisme o evolucionar cap a formes verbals menys dramàtiques, com l'anàrtria o l'agramatisme.
2. **Anàrtria**: presència de problemes fonètics en forma d'omissions, addicions i substitucions (DIÉGUEZ i PEÑA, 2012: 88), que provoquen dificultats en l'emissió verbal i un conjunt de disfuncions articulatòries (DIÉGUEZ i PEÑA, 2012: 149).
3. **Agramatisme**: problemes en la producció de morfemes gramaticals, que pot portar a la producció d'una parla telegràfica (DIÉGUEZ i PEÑA, 2012: 88). També es pot donar per alteracions de l'ordre sintàctic o disfuncions prosòdiques (DIÉGUEZ i PEÑA, 2012: 168). Una parla agramàtica sol estar formada per oracions declaratives, sense clàusules de relatiu, adjunts en forma de sintagmes preposicionals o especificadors nominals o verbals (DIÉGUEZ i PEÑA, 2012: 185).
4. **Parafàsia**: substitució d'un element per un altre de la mateixa categoria. Poden existir diversos tipus de parafàsies, com les verbals, les fonètiques o les semàntiques. Els seus mecanismes responsables són la substitució, l'addició, l'omissió i la distorsió (DIÉGUEZ i PEÑA, 2012: 145).
5. **Denominació**: acte lingüístic que consisteix en la cerca i posterior emissió d'un element, que normalment és una paraula.
6. **Latència**: cerca d'una paraula, que pot executar-se sense dir res o amb conductes verbals del tipus "no ho sé" o amb sons com "um" o "ah".
7. **Paraules òmnibus**: paraules buides de significat, generals, com *això, aquell, cosa, objecte, aquí, res*, etc.

7.3. *Annex III:*

PROTOCOL DE TRANSCRIPCIÓ

PASSOS

1. Transcripció automàtica amb el programa Buzz
2. Revisió i modificació (si escau) de la transcripció generada automàticament
3. Guardar amb format .srt
4. Transformar el text de .srt a TextGrid
5. A Praat, generar el nivell de “pausa i parla”.
6. Combinar el TextGrid de “pausa i parla” amb el de la transcripció
7. Obrir el document generat juntament amb l’àudio
8. Crear els 5 nivells (Tiers)

TIERS

- Tier1: transcripció paraula
    - Transcripció paraula
    - Sense puntuació
- tier 2:
    - Transcripció nivell respiratori
    - Dins una mateixa emissió amb pauses internes: Primer límit quan comença el silenci, últim límit quan acaba la parla. (igual que entre interlocutors)
    - Etiquetes (quan no es transcrigui la parla del pacient):
        - *interlocutor*
        - *inintel·libible*
- Tier 3: parla-pausa
    - Pausa i Parla
- Tier 4: comentaris
    - Comentaris potencialment útils per a l’anàlisi
    - Etiquetes:
        - Pausa plena
        - Sons de l’estil “mmm”, “eehhh” o muletilles (ex:”Bueno”)
- Tier 5: comentaris extra
    - Comentaris extra que poden ajudar a un anàlisi posterior

- Etiquetes:
    - Canvi de codi: emissió en la llengua no objectiu
    - Insta canvi de codi: intervenció de l'interlocutor que recorda la llengua de la prova
    - Interferència:
    - Interferència acceptada: terme que una persona sana del mateix entorn sociolingüístic utilitzaria
    - Pacient: utilitzat per indicar que és el torn de parla del pacient, o que és qui fa la pausa o pausa plena
    - Titubeig:
    - No transcrit, inclòs en pausa plena
    - Autocorrecció
    - Allargament
    - Anotat a partir de (>200ms)
    - Repetició
    - Solapament
    - En algunes ocasions, possible interpretació de sons a priori *inintel·ligibles*

PAUSES

- Pauses interconversacionals – comencen a l'interlocutor 2 quan l'interlocutor 1 ha acabat de parlar
- Etiquetades a partir de 150ms

INTERFERÈNCIES LINGÜÍSTIQUES

- Barbarismes, manlleus i canvis de codi - es transcriuen sense traduir
- Interferències lingüístiques i canvis de codi: etiquetar-ho al nivell 5

ALTRES

- Parla inintel·ligible:
    - Només s'etiqueten com a pauses els silencis, no les pauses plenes, perquè és difícil diferenciar què és titubeig, fals inici... Per tant, emissió=parla, silenci =pausa
- Solapament:

    - Si és intel·ligible, es prioritza la transcripció del pacient als tiers 1 i 2. S'anota "solapament" al tier 5.
- Parafàsies i erros de pronunciació: transcriure la paraula objectiu
- Elisions
- Vocàliques o consonàntiques: es transcriuen segons l'ortografia de la llengua. Exemple: [me] □ "me he"
- Resta: es transcriu el que pronuncia el pacient. Exemple: en català, a "no sé", no s'hi afegeix "ho" (gramaticalment correcte) si el pacient no ho pronuncia.