UNIVERSITAT DE BARCELONA

Màster d'Humanitats Digitals

Curs 2020 – 2021

Segon semestre

Anotació del focus de la negació i de la temporalitat en informes mèdics

Autora: Laura Tañá Velasco

NIUB: 16618630

Tutora acadèmica: Mariona Taulé Delor

Data: setembre 2021

Agraïments

Voldria agrair a la Mariona Taulé, la M. Antònia Martí i la Montserrat Nofre l'assessorament, la revisió i la implicació al llarg de tot el procés d'elaboració del treball. Sense elles no hauria estat possible.

Resum

El volum de la documentació que es genera en l'àmbit de la pràctica mèdica creix exponencialment. Les dades clíniques sovint es troben no estructurades o semiestructurades, pel que cal aplicar tècniques avançades d'extracció d'informació per millorar l'accés, l'anàlisi i la interoperabilitat de les dades dels pacients. Una d'aquestes tècniques és l'Aprenentatge Automàtic, la qual permet entrenar models que extreguin informació de manera automàtica o supervisada a partir de corpus anotats.

Per a l'aplicació d'aquestes tècniques cal afrontar l'ambigüitat inherent del llenguatge natural i, en el cas de la documentació mèdica, els desafiaments que presenta el domini mèdic atès que es tracta d'un subllenguatge amb característiques pragmàtico-semàntiques, lèxiques, sintàctiques i ortotipogràfiques específiques.

En aquest treball, *Anotació del focus de la negació i de la temporalitat en el domini mèdic*, presentem les característiques del subllenguatge mèdic i ens centrem en el tractament del focus de la negació en documents del domini mèdic per a l'ensinistrament de sistemes de detecció de la negació basats en l'Aprenentatge Automàtic. En l'àrea de l'extracció d'informació l'expressió de la negació encara resulta un aspecte problemàtic, tot i que el seu tractament és important per comprendre correctament els textos. Volem contribuir en l'estudi del focus de la negació i crear un nou recurs lingüístic, el corpus *ClUB-21* i la guia d'anotació corresponent. Tractem també la temporalitat i els diferents tipus d'expressions temporals per l'ambigüitat que generen a l'hora d'identificar el focus de la negació.

Paraules clau

Extracció d'informació, Aprenentatge Automàtic, informes mèdics, focus de la negació, negació, temporalitat, corpus, guia d'anotació.

Índex

1. Introducció 4
2. Conceptes bàsics 6
2.1. *La negació i l'especulació* 6
2.2. *Expressió del temps* 8
2.3. *Característiques del subllenguatge mèdic* 10
2.4. *Anotació de corpus per a l'Aprenentatge Automàtic* 16
3. Objectius 18
4. Estat de l'art 19
4.1. *Estat de l'art de la negació* 19
4.2. *Estat de l'art: expressions temporals* 26
5. Metodologia 33
5.1. *El corpus ClUB-21* 33
5.2. *La negació en el domini mèdic* 36
5.2.1. *Tipus d'estructures negatives en el domini mèdic* 37
5.2.2. *Problemes específics de la negació en el domini mèdic* 39
5.2.3. *El focus de la negació* 41
5.3. *Expressions temporals en el domini mèdic* 45
5.3.1. *Problemes específics de les expressions temporals en el domini mèdic* 49
6. Proposta d'anotació del focus de la negació en el domini mèdic 53
7. Línies de futur 60
8. Conclusions 61
9. Apèndix 62
9.1. *Llistat de marcadors de negació del domini mèdic* 62
10. Bibliografia i referències 64

1. Introducció

El projecte de recerca que presentem s'emmarca en el camp de les tecnologies del llenguatge i, més concretament, en l'extracció d'informació en el domini mèdic. L'extracció d'informació és un procés automàtic per obtenir informació a partir de documents en llenguatge natural en format digital. Els documents poden tenir una gran varietat de formats: estructurats, semiestructurats i no estructurats (text lliure). Normalment l'extracció d'informació s'aplica a documents no estructurats en llenguatge natural o a documents semiestructurats en llenguatge natural.

En l'àmbit de la pràctica mèdica, especialment en els grans centres assistencials, la informació que es genera es troba ja en format digital. Aquesta informació creix de manera exponencial. El contingut dels documents generats (informes d'alta, ingressos, atenció d'urgències, proves clíniques, etc.) seria de gran valor si estès disponible de manera estructurada en un format de base de dades. El problema és que en l'actualitat els sistemes d'extracció d'informació no són capaços de processar els documents i convertir les dades en informació i coneixement utilitzable. Cal desenvolupar tècniques avançades en l'extracció d'informació per fer accessibles aquests continguts. Com a resultat, es podrà oferir una millor atenció als pacients, una millora en la gestió de la informació en els grans centres hospitalaris i serà possible la recerca de qualitat.

Algunes de les principals capacitats que han de complir els sistemes d'extracció d'informació segons Nadeau i Sekine (2007), entre d'altres, són:

> **El reconeixement d'entitats nombrades**. Cada domini de coneixement té les seves pròpies entitats, tot i que n'hi pot haver de comunes a diferents dominis, com ara els noms de persona, de lloc o les dates. En el domini mèdic es tracten entitats com malalties, tractaments, proves mèdiques, medicaments o símptomes. Serien exemples d'aquesta terminologia: 'càncer', 'radioteràpia', 'anàlisi d'orina', 'febre', etc.
>
> **L'extracció de relacions**. L'extracció de relacions consisteix a detectar les relacions semàntiques entre entitats. En el domini mèdic cal extreure relacions del tipus: malaltia-malalt, malaltia-medicament, medicament-temps, etc.
>
> **L'anàlisi de la correferència**. L'anàlisi de la correferència consisteix a determinar les expressions lingüístiques que fan referència a la mateixa entitat del món real. Consisteix, per tant, a identificar, per exemple, que 'OD' és el mateix que 'ojo derecho'.
>
> **La definició d'esdeveniments** com són els ingressos o les proves mèdiques.

D'acord amb estudis com el de Ben Abacha i Zweigenbaum (2011), el volum del coneixement mèdic es dobla cada cinc anys. Les dades clíniques es troben en forma de documents mèdics escrits per professionals de la medicina, són de difícil accés i sovint no estructurades. És necessari, doncs, aplicar tècniques avançades d'extracció d'informació per millorar l'accés, l'anàlisi i la interoperabilitat de les dades dels pacients. El subllenguatge mèdic, a més, es

caracteritza per tenir peculiaritats lingüístiques pròpies i, per tant, necessita d'un tractament especialitzat.

A banda de les dificultats que presenta el Processament del Llenguatge Natural a causa de, per exemple, l'ambigüitat inherent del propi llenguatge, el domini mèdic, a més, presenta desafiaments propis degut a les seves característiques pragmàtico-semàntiques, lèxiques, sintàctiques i ortotipogràfiques específiques. Ho tractarem amb més detall a l'apartat 2.3.

En aquest treball ens centrarem en la problemàtica que planteja l'expressió de la negació i de la temporalitat per ser dues qüestions rellevants en els sistemes d'extracció d'informació del domini mèdic. D'una banda, la negació modifica el valor de veritat d'una declaració i, per tant, és necessari el seu tractament per a la correcta comprensió d'un text. Ens centrarem concretament en la identificació del focus de la negació ja que és l'element negat més important. D'altra banda, tractarem la temporalitat per l'ambigüitat que genera a l'hora d'identificar el focus de la negació, així com la identificació de les relacions temporals entre esdeveniments clínics.

L'objectiu concret d'aquest treball és definir els criteris per a l'anotació del focus de la negació i les expressions temporals en el corpus *ClUB-21*, un corpus de documents mèdics, per tal que serveixi com a corpus d'entrenament per a sistemes de detecció automàtica de la negació i de les expressions temporals.

L'estructura del treball és la següent:

- Definició dels conceptes bàsics del nostre treball (apartat 2), els quals són:
    - La negació i l'especulació (2.1.).
    - Les expressions temporals (2.2.).
    - Les característiques del subllenguatge mèdic (2.3.).
    - L'anotació de corpus per a l'Aprenentatge Automàtic (2.4.).
- Objectius del nostre treball (apartat 3).
- L'estat de l'art de la negació i de la temporalitat en el domini mèdic (apartat 4).
- Metodologia que hem seguit (apartat 5):
    - Les característiques del nostre corpus, el *ClUB-21* (5.1.).
    - La problemàtica que planteja la negació a partir d'exemples del domini mèdic (5.2.).
    - La problemàtica que planteja la temporalitat a partir d'exemples del domini mèdic (5.3.).
- Una proposta d'anotació del focus de la negació en el domini mèdic (apartat 6).
- Línies de futur del nostre treball (apartat 7).
- Conclusions (apartat 8).

2. Conceptes bàsics

En aquest apartat definim aquells conceptes que són bàsics per al nostre projecte i que, per tant, són a la base del treball realitzat. Això és: l'expressió de la negació i de la temporalitat ja que són els objectes del nostre estudi. A més, incloem el tractament de l'especulació atès que la negació hi està relacionada i sovint es confon amb ella. Presentem també característiques del subllenguatge mèdic ja que és un domini concret i, per tant, presenta característiques pròpies que afecten l'expressió de la negació i de la temporalitat. Finalment, expliquem en què consisteix l'anotació de corpus per a l'Aprenentatge Automàtic ja que és dins d'aquest marc on es situa el nostre projecte.

2.1. La negació i l'especulació

La **negació** és un fenomen compartit probablement a totes les llengües i que consisteix a modificar el valor de veritat d'una declaració o d'un esdeveniment o a fer explícita l'absència d'un esdeveniment.

Els sistemes de detecció de la negació focalitzen el problema en la detecció dels components següents:

- Detecció del **marcador de negació**. El marcador és aquella paraula o grup de paraules que identifica que estem davant d'un fet negat. Normalment els marcadors són de nombre finit i molts sistemes de detecció automàtica de la negació els incorporen en forma de llistes. En general pertanyen a categories tancades com són els adverbis i els pronoms, però cal tenir en compte que també hi ha marcadors de base lèxica i, en aquest cas, la llista és oberta i no han de per què contenir cap expressió de negació. Per exemple, l'expressió lèxica 'en la vida' funciona com un adverbi i és equivalent al marcador adverbial 'nunca', però no conté cap indici extern d'expressió de la negació. Com que el marcador és l'element que dona lloc a la negació, la seva detecció sol ser la primera tasca en els sistemes de processament de la negació. Són exemples de marcadors de negació: 'no', 'ni', 'mai', 'ningú', 'finalitzar', 'negatiu', etc.

- Identificació de **l'abast** (*scope,* en anglès). L'abast es sol definir com la màxima unitat sintàctica que es troba afectada pel marcador de negació (Morante i Sporleder, 2012). Està conformat, per tant, pels elements que es veuen afectats per la negació. Es tracta d'una definició vaga que s'interpreta de manera molt lliure. Això origina que estigui representat de diferents maneres als corpus: per exemple, en algunes anotacions s'inclou dins de l'abast el marcador de negació, però en altres no. A l'oració "No [signos de irritación peritoneal]", els elements entre claudàtors són l'abast de la negació.

- Identificació de **l'esdeveniment** (*event*, en anglès) que es nega. Aquest concepte planteja problemes semblants als de l'abast, però en relació amb els fets que es consideren com a esdeveniments. Normalment està associat a verbs i en molts casos no s'anota. En el nostre cas, no l'hem tingut en compte. A l'oració "No *se hizo* la prueba de SARS-CoV-2", l'element en cursiva és l'esdeveniment.

- Modificació de **la polaritat** (*polarity*, en anglés). S'entén com a polaritat aquells elements que reforcen la negació i que incrementen o disminueixen el grau de negació. Generalment són adverbis, pronoms o adjectius. Per exemple, a l'oració "**No** tomaba ninguna medicación", 'ninguna' seria un element que expressaria polaritat i reforça la negació 'no'.

- Detecció del **focus de la negació**, és a dir, allò que és específicament negat. El focus es sol definir com la part de l'abast més explícita o prominentment negada (Huddleston i Pullum, 2002). A l'oració "No reveló ninguna arteriopatía coronaria obstructiva", 'ninguna arteriopatía coronaria obstructuva' és el focus de la negació.

D'acord amb Jiménez-Zafra et al. (2019), hi ha diferents tipus de negació segons el marcador que s'utilitza:

- **Negació sintàctica**. La negació sintàctica es tracta de la negació expressada per un marcador, normalment una paraula funcional: una preposició, una conjunció o un adverbi. Les oracions següents són alguns exemples[1] de negació sintàctica que trobem al nostre corpus:
  (1)
  a. Deposiciones explosivas liquidas[2] de color claro **sin** productos patologicos
  b. **No** DM, **ni** cardiopatia, **ni** tumores previos
  c. **No** signos de irritación peritoneal

- **Negació lèxica**. La negació lèxica es troba expressada mitjançant marcadors de base lèxica com són els verbs, els substantius i els adjectius. A continuació presentem alguns exemples de negació lèxica del nostre corpus:
  (2)
  a. La presencia de eosinófilos obligan a **descartar** un componente de colitis farmacològica
  b. **Falta de** medicació
  c. En el día de ayer PCR COVID positivo compañero de habitación, por lo que se realitza PCR al mismo siendo resultado **negativo**

[1] Tots els exemples que apareixen en aquest treball estan extrets del corpus que hem creat, el *ClUB-21*.
[2] En els exemples es manté la literalitat original tant si hi ha errors ortogràfics com tipogràfics.

- **Negació morfològica**. La negació morfològica s'expressa mitjançant afixos, és a dir, sufixos o prefixos. Alguns exemples de negació morfològica que trobem al nostre corpus són els següents:
  (3)
  a. **A**febril
  b. HDM estable, HTA **a**sintomático, **a**febril

L'**especulació o incertesa** és un fenomen que té lloc quan un parlant expressa de forma explícita que no està segur del que es diu o que no sap del cert alguna cosa. És un tipus d'expressió freqüent en el domini mèdic ja que és una manera habitual d'expressar alertes o conjectures sobre l'estat del pacient o sobre les dades de les proves que se li han aplicat. Es tracta, doncs, d'una informació d'especial rellevància pel domini mèdic en comparació amb altres subllenguatges ja que la detecció de fets presentats com a hipotètics o possibles pot tenir repercussions importants de cara a l'evolució i el tractament dels pacients. Són exemples d'expressió de l'especulació:

(4)
a. Sospecha clínia de síndrome coronario agudo
b. Posibilidad de infección por SARS-CoV-2
c. Parecía tener una leve congestión

2.2. Expressió del temps

La **informació temporal** permet situar un esdeveniment en una línia temporal. Aquesta informació s'expressa mitjançant estructures o paraules que es refereixen al moment, a la durada o a la freqüència d'un esdeveniment.

En el domini mèdic són freqüents les estructures temporals ja que es fa referència a entitats o esdeveniments, com poden ser els símptomes, els tractaments o els procediments, que s'esdevenen o no només en un lapse determinat de temps. És important col·locar aquests esdeveniments i aquestes entitats en un context temporal per entendre l'ordre cronològic dels procediments clínics i dels tractaments.

Per al tractament de la temporalitat és rellevant l'anotació de diferents tipus d'informació:

- **Els esdeveniments**. D'acord amb Pustejovsky et al. (2003) els esdeveniments poden definir-se com qualsevol cosa sobre la qual es pugui dir que ha estat obtinguda, que és certa, que ha succeït o que ha tingut lloc. L'objectiu de la detecció de la informació temporal és precisament ancorar els esdeveniments en un marc temporal per entendre el seu ordre cronològic.

- **Les estructures temporals**. Es tracta d'expressions lingüístiques que donen informació sobre quan succeeix un esdeveniment, la seva durada o bé la freqüència amb què succeeix.
- **La relació temporal** entre dos esdeveniments, entre una expressió temporal i un esdeveniment o bé entre dues expressions temporals. És a dir, si un esdeveniment, per exemple, és anterior, posterior, simultani, etc. respecte d'un altre esdeveniment.

Analitzem les següents oracions:

(5)

a. **El dolor va augmentar** el dia abans de **l'ingrés**
b. **Remisión de la fiebre** a las 30 horas de **hospitalización**

En l'exemple (5a) s'expressen dos esdeveniments que són 'el dolor va augmentar' i 'l'ingrés' i l'estructura temporal 'el dia abans' que estableix la relació entre els dos esdeveniments ja que especifica que el primer és anterior al segon. A més, 'el dolor va augmentar' és simultani a l'estructura temporal, però aquesta és anterior a 'l'ingés'.

En l'exemple (5b) trobem també dos esdeveniments que són 'remisión de la fiebre' i 'hospitalización'. L'estructura temporal és 'a las 30 horas' i és anterior a 'remisión de la fiebre' i simultània a 'hospitalización'. A més, el primer esdeveniment és posterior al segon.

Cal assenyalar, a més, que aquestes expressions poden tenir diferents valors: poden fer referència a un període o interval de temps, a una iteració de processos o bé a una data o hora puntual.

Les següents oracions són exemples d'estructures temporals puntuals, és a dir, que fan referència a un moment puntual del calendari. Es solen expressar mitjançant dates o hores, pel que poden contenir paraules temporals com 'dia', 'mes', 'any' o 'hora', com veiem en l'exemple (6b):

(6)

a. Paciente de 85 años que acude a Urgencias el 12.01.2021 tras presentar caída casual sin TCE
b. Ingressa el dia15.03.2021
c. 23.02.2021 23h: 38ºC T axilar

Els següents exemples són estructures temporals que indiquen la duració d'un procés i que, per tant, fan referència a un període en què succeeix l'esdeveniment. Al marcar un interval de temps tenen un inici i un final, però que no són explícits a vegades, pel que no podem conèixer-los, com observem en l'exemple (7a):

(7)

a. Desde hace unos cuatro días se ha iniciado también escasatos
b. Durante la tarde comenta que siente frío y sufre una tiritona
c. Aquest pacient ha tingut contacte amb un cas confirmat de covid19 entre els dies 11/12/2020 fins el 14/12/2020

Finalment, les següents oracions contenen estructures temporals que marquen una iteració de processos, és a dir, una freqüència. Fan referència a esdeveniments que es repeteixen en el temps:

(8)

a. Pantoprazol 20 MG cada 24 horas
b. Diuresis 1450 cc/24h
c. Recomienda control cuentas CMV cada 15 días para valorar evolución

2.3. Característiques del subllenguatge mèdic

En aquest apartat presentem les característiques del subllenguatge mèdic. Partim de dos tipus d'informació. D'una banda, s'ha dut a terme un estudi exhaustiu del corpus que ens ha cedit l'Hospital Clínic de Barcelona, el corpus *ClUB-21*, i d'on procedeixen els exemples. D'altra banda, incorporem el treball realitzat per la Dr. Estopà (2020) sobre el subllenguatge mèdic.

Segons aquesta autora "un informe mèdic és un text escrit per un metge sobre el procés assistencial d'un pacient, en el qual es descriuen processos, proves i observacions per tal d'arribar a un diagnòstic i a un tractament adequat". Per tant, té com a objectiu registrar les dades necessàries per assistir mèdicament al pacient i presenta diferents funcionalitats: legal, científica, comunicativa, biogràfica i d'avaluació. Es tracta, doncs, d'un text especialitzat que conforma un gènere textual propi.

Pel fet de tractar-se de textos d'un domini específic, presenten característiques lingüístiques, tant pragmàtico-semàntiques com lèxiques, sintàctiques i ortotipogràfiques, pròpies.

Cal tenir en compte també les situacions comunicatives en què es produeixen els informes mèdics. L'informe es redacta o bé al final d'una visita o bé al llarg de diferents interaccions amb el pacient que es produeixen amb interrupcions i, tant en un cas com en l'altre, el temps de redacció del que disposa el facultatiu és limitat. L'espai de redacció també presenta limitacions i generalment els centres hospitalaris ofereixen al personal unes plantilles o bé en paper o bé electròniques que estructuren el contingut. Tot això condiciona la forma en què s'expressa la informació que es recull.

Cal afegir també el fet que aquests textos estan pensats per l'ús intern i, per tant, destinats a altres especialistes.

Tenint en compte aquests condicionaments generals, presentem a continuació les característiques que hem observat en els documents del nostre corpus.

a) Expressió de les formes nominals

Podem observar que en els informes mèdics predominen les expressions nominals i l'absència de formes verbals. Observem un predomini de substantius i d'elements associats al sintagma

nominal com són els adjectius o els sintagmes preposicionals complementant un nom en els següents exemples:

(9)

a. Progresiva disminución de volumen urinario, y limitación para la movilización
b. Micciones abundantes en BGP
c. Paciente consciente y orientado, con tendencia a la HTA y afebril

Trobem també una gran abundància de termes del domini en relació amb el número total de paraules. L'ús de terminologia especialitzada permet comunicar de manera precisa i adequada coneixements de l'àmbit mèdic. Tenen, doncs, una abundància de tecnicismes i de noms propis com es pot observar en els següents exemples:

(10)

a. ECG se observa ritmo sinusual a 75 lpm, QRS estrecho, sin alteraciones en conducción ni repolarización
b. En mi valoración con TA: 130/78 mmHg, FC: 70lpm, SatO2: 98% basal
c. Solicito A/S, frotis y Rx tórax de control para miércoles

b) Expressió de les formes verbals

En el subllenguatge mèdic és freqüent l'elisió de verbs. Aquest tret és degut a que aquest tipus de documents es produeixen en un context en què el facultatiu disposa de poc temps i l'el·lipsi permet una redacció més ràpida. Tanmateix això no dificulta la comprensió ja que els verbs elidits solen ser verbs de significat existencial (presentar, haver-hi, ser, etc.). Les frases següents són alguns exemples de l'elisió de verbs:

(11)

a. No signos de irritación peritoneal
b. Habla fluente y discurso coherente
c. No focalidad neurológica aguda
d. Paciente de 67 años, sin alergias medicamentosas, no hábitos tóxicos

Les formes verbals que apareixen estan sovint en forma no personal, és a dir, es tracta d'infinitius, gerundis i participis.
S'observa també un ús escàs d'oracions subordinades atès que es busca una redacció ràpida i, per tant, es prefereixen oracions simples.
Predomina l'ús de l'infinitiu per expressar una ordre, una instrucció o una recomanació i del gerundi i del participi amb un matís adverbial en lloc d'oracions compostes subordinades. En podem observar alguns exemples en les oracions següents:

(12)

a. Transfundir 2 concentrados de hematíes hoy

b. Se aplicó un tratamiento sintomático con ceftriaxona i.v., teniendo en cuenta su leucocitopenia y el alto riesgo de infección urinaria, pero se retiró después de que los hemocultivos y los urocultivos, obtenidos antes de la administración de antibiótico, no mostraran crecimiento a las 24 horas
c. Función renal normal con diuresis preservada
d. Peristaltismo positivo, aumentado

Predomina l'ús de verbs en forma impersonal, construccions passives i passives reflexes. Aquests recursos permeten evitar la subjectivitat i reforçar l'objectivitat científica. És sobretot freqüent l'ús d'aquest tipus de construccions en les oracions que es refereixen a accions o decisions mèdiques. En són un exemple:

(13)
a. Es dóna mediació immusupressora per 24h i es dóna l'alta
b. Se ha tratado poco tiempo la colitis
c. Se decide traslado a nuestro centro por evidencia de desaturación a 90% basal con lentillas nasales a 3 litros

c) Parataxi i hipotaxi

La parataxi i la hipotaxi són diferents tipus de relacions gramaticals que s'estableixen entre els diferents constituents sintàctics: oracions, sintagmes o paraules. La parataxi consisteix en la unió de dos o més elements sintàctics o oracionals mitjançant la coordinació o la juxtaposició. D'altra banda, la hipotaxi consisteix en la unió de dues o més oracions subordinades a una de principal.

En relació amb aquests procediments sintàctics, en els informes mèdics observem les següents característiques:

Aquests documents presenten poca hipotaxi o subordinació, és a dir, poca presència d'oracions compostes subordinades. Això és a causa de la simplicitat sintàctica que els caracteritza, la qual està afavorida per les limitacions de temps i d'espai de què disposen els facultatius a l'hora de redactar. S'elimina l'oració principal que té un caràcter introductori i s'expressa directament la subordinada, on es troba el focus de la informació.
Les següents oracions són exemples on no trobem subordinació:

(14)
a. No algies ni malestar
   Sense incidències
   Descans nocturn

b. Se inicia en fst respiratoria
   Paciente ligeramente nerviosa
   Deambula de manera autónoma
c. Pacient autònom
   Sedesta i deambula per interior habitació

Tanmateix a vegades sí que trobem oracions subordinades quan en la oració principal s'expressa informació imprescindible. Les oracions subordinades introduïdes per una oració principal en la seva major part són:

- Oracions subordinades de relatiu, atès que en aquests casos és imprescindible la presència de l'antecedent:
  (15)
  Se colocó un tubo de drenaje torácico, que posteriormente se retiro

- Oracions subordinades temporals, directament relacionades amb l'expressió de la temporalitat i que aporten informació imprescindible per entendre l'ordre dels procediments:
  (16)
  Se le aplicó un tratamiento sintomático con ceftriaxona i.v., teniendo en cuenta su leucocitopenia y el alto riesgo de infección urinària, pero se retiró después de que los hemocultivos y los urocultivos, obtenidos antes de la asministración del antibiótico, no mostraran crecimiento a las 24 horas

Hi ha, a més, un domini de la parataxi entre els elements sintàctics i les oracions. És freqüent, per exemple, la parataxi en l'exposició de símptomes, com s'observa en els següents exemples:

(17)

a. Al ingreso, la paciente tenía fiebre de 5 días, conjuntivitis sin afectación limbal, papilas gustativas prominentes, exantema maculopapuloso pálido poli mórfico e hinchazón de las manos y las extremidades inferiores, manifestaciones coherentes con los criterios clásicos de la enfermedad de Kawasaki
b. A su llegada a urgencias, hemodinámicamente estable, afebril y eupenica
c. No mostraba tos, congestión ni rinorrea

També és freqüent la parataxi en forma de llistes. Les llistes són la màxima expressió de la simplicitat sintàctica i s'utilitzen per aconseguir una exposició clara i concisa. Alguns exemples de llistes que trobem al nostre corpus són:

(18)

a. Se realizan pruebas complementarias: Analítica: PCR 2.63, FG 73.68 ml/min, AST 14, BT 0.30, FA 52/GGT 15, Leucopenia 3650 (No linfopenia), Hb 13.6, Plaquetas 157000, dimero D 1500. Radiografía de tórax: Se observan infiltrados intersticiales periféricas bilaterales de predominio en las bases. Electrocardiograma: Ritmo sinusal con FC, no alteración en PR, QTc 359 mseg, no cambios en la repolarización (Fallo de electrocardiografo, no realiza V5)
b. PLAN: Mantenemos tratamiento empírico con Hidroxicloroquina + Azitromicina+Lopinavir/Ritonavir. Pendiente de resultados de Hemocultivos

30.03. Mantener oxigenoterapia (gafas nasales 2L). Solicitamos analítica en 48h para 03.04. Solicitamos Rx tórax 02.04

c. **ANTECEDENTES PATOLÓGICOS: HIPERTENSIÓN ARTERIAL ESENCIAL desde hace más 5 años. En seguimiento en H. Clinic. DIABETIS MELLITUS tipo 2, diagnosticada en 2018 en tratamiento con insulina. Última HbA1c 6.4% 10/19. Sin afectación de órgano diana. HERNIA DE HIATO pequeña, con fibrogastroscopia en 2017 con pólipos gástricos con AP benigna. INSOMNIO DE CONCILIACIÓN

d) Abundància d'expressions temporals

Els informes mèdics presenten una gran abundància d'estructures temporals ja que els esdeveniments sempre van lligats a un lapse de temps. És important col·locar-los en un context temporal per entendre l'orde cronològic dels successos que s'exposen. Com es veurà més endavant (apartat 5.3.), però, un dels problemes que plantegen les expressions temporals és saber a quins esdeveniments s'han d'aplicar i què és allò que està sent delimitat temporalment.

Cal assenyalar, a més, que aquestes expressions tenen diferents valors: poden fer referència a un període o interval de temps, a una iteració de processos, a una data o hora puntual, etc. Hi ha casos, com 'actualmente', en què les estructures temporals són de temps relatiu i que, per tant, és difícil situar l'esdeveniment al qual es refereixen en una línia temporal.

(19)

a. Ultima deposicion el dia 30/11 a las 13h
b. A las 3h refiere dolor de espalda, administro paracetamol, es efectivo
c. Mastical 500 mg 1 comprimido cada 12 horas
d. Última depo 30/11/20. Actualmente con diarrees

e) Estructures de valor-resultat

Trobem també la presència d'apartats on s'exposen proves o símptomes en forma d'estructures de valor-resultat. Sovint aquestes estructures són predeterminades. Això permet també una redacció més ràpida i una exposició més clara i concisa. Els exemples següents són alguns fragments del nostre corpus amb estructures de valor-resultat:

(20)

a. Tipus de respiració: Eupneic
   Valor Freqüència respiratòria: 20 resp/min
   Valor Pressió Arterial sistòlica: 119 mm HgValor
   Pressió Arterial diastòlica: 76 mm Hg
   Valor Freqüència cardíaca: 112 lpm
b. Diarrea: Si
   Pròtesi dental: Si
   Alteració de la integritat de la pell a l'ingrés: No
   Coloració de la pell: Normocoloreada
   Puntuació escala Braden: 22-Sense Risc

Temperatura axil·lar: 34 ºC
Estat de la consciència: Conscient
Té dolor?: No

f) Especulació

Observem, a més, una presència abundant d'expressions hipotètiques o especulatives. Aquest tipus d'expressions es donen quan el facultatiu no està segur d'una deducció i així evita ser categòric. Són expressions que indiquen, per tant, incertesa o especulació i estan introduïdes, principalment, mitjançant paraules que indiquen possibilitat o sospita o bé per verbs com 'suggerir' o 'semblar'. Alguns exemples d'expressions hipotètiques són:

(21)
a. La radiografía torácica señalaba la posibilidad de infección por SARS-CoV-2
b. Sospecha clínica de síndrome coronario agudo
c. Parecía tenir una leve congestión
d. Una radiografía torácica mostró signos de neumonía intersticial grave, con alteraciones en vidrio esmerilado, que sugerían una infección por SARS-CoV-2

g) Lèxic idiosincràtic del domini mèdic

El subllenguatge mèdic es caracteritza també per l'ús d'expressions i verbs específics del domini mèdic que rarament es troben en textos que pertanyen a altres dominis. Alguns exemples extrets del nostre corpus són els següents:

(22)
a. Sedesta en silla y va alternando la cama
b. Ingesta correcta
c. Deambula de manera autònoma amb LN + allargadera
d. Habla fluente y discurso coherente

g) Presència de més d'una llengua

Els informes mèdics són el resultat de la intervenció de molts facultatius. Com a resultat, en el corpus *ClUB-21*, ens trobem amb què els textos es troben escrits en més d'una llengua. A Catalunya, a més, el bilingüisme català/espanyol fa que aquestes dues llengües convisquin en els informes. Els informes de curs clínic, com que es van elaborant al llarg de tota l'estada del pacient al centre mèdic, estan redactats per diferents professionals, pel que és normal que un mateix informe contingui parts en diferents idiomes i diferents estils de redacció. A més, en el cas dels informes d'alta es reaprofiten parts d'altres documents, la qual cosa afavoreix que continguin llengües i estils diferents.

Aquestes característiques lingüístiques pròpies de la documentació que es genera en l'entorn mèdic fan necessària l'aplicació de criteris específics per al tractament de la negació i del seu

focus i per al tractament de la temporalitat ja que no sempre poden aplicar-se criteris propis de la llengua estàndard.

2.4. Anotació de corpus per a l'Aprenentatge Automàtic

L'Aprenentatge Automàtic és una tècnica desenvolupada en el marc de la intel·ligència artificial que té com a objectiu desenvolupar programes que permetin aprendre a partir de dades o d'exemples. Concretament, es tracta de crear algoritmes capaços de generar comportaments i de reconèixer i classificar patrons a partir de mostres (dades o exemples) elaborades específicament per a aquest fi. En el Processament del Llenguatge Natural s'utilitzen tècniques d'Aprenentatge Automàtic per entrenar models que extreguin informació de manera automàtica o supervisada a partir de corpus anotats. Per a l'entrenament d'aquests models automàtics calen corpus o conjunts de textos anotats de manera manual per humans. A partir, doncs, d'aquestes dades etiquetades els ordinadors aprenen a analitzar el llenguatge natural i a identificar patrons.

L'Aprenentatge Automàtic és la tècnica sobre la qual més s'investiga actualment. És una aproximació al tractament automàtic del llenguatge basada en mètodes estadístics. Es tracta, per tant, d'una aproximació *bottom-up*, és a dir, d'una aproximació en què a partir de les dades s'infereix el coneixement.

En l'actualitat, els corpus anotats són un recurs lingüístic bàsic per desenvolupar aplicacions del Processament del Llenguatge Natural. Consisteixen en una col·lecció de textos escrits o orals creada amb una finalitat concreta i a partir d'un conjunt de criteris ben definits. La seva anotació consisteix a afegir informació lingüística mitjançant etiquetes per fer més ràpida i fàcil l'obtenció i l'anàlisi de la informació. Els corpus s'han d'anotar amb la informació que volem que aprengui el sistema d'Aprenentatge Automàtic. Si anotem el corpus amb informació morfològica, el programa d'aprenentatge aprendrà a anotar corpus amb aquesta informació. En el cas que ens ocupa, anotem el corpus amb informació sobre la negació i les expressions temporals per tal que el sistema sigui capaç de detectar aquesta informació en un corpus no anotat.

En funció de l'estudi que es realitza, s'han de definir els criteris per a l'anotació. Cal elaborar, en primer lloc, una guia d'anotació en què s'especifiqui l'objectiu de la tasca, es defineixi l'etiquetari i s'expliquin els criteris. És important també determinar el perfil dels anotadors. Aquesta fase és iterativa ja que la guia es va revisant i modificant progressivament fins que s'assoleix un acord acceptable en el treball d'anotació. És una fase d'entrenament que serveix per elaborar la guia d'anotació definitiva i per a l'ensinistrament dels anotadors.

La qualitat del corpus anotat es mesura en termes de l'*Inter-Annotator Agreement* (IAA) o proves d'acord entre anotadors. Per a dur a terme les proves d'acord entre anotadors cal seguir el procediment següent:

a. En primer lloc, cal determinar què és el que es vol anotar i quins materials hi ha disponibles sobre el tema. És important basar-se en treballs realitzats prèviament per altres investigadors i avaluar els aspectes positius i les mancances que presenten.
b. A partir d'aquest primer estudi s'elabora una guia d'anotació provisional que servirà com a punt de partida per al procés d'anotació.
c. Es seleccionen de manera aleatòria uns fragments del corpus per dur a terme una anotació en paral·lel. L'ideal és que l'anotació en paral·lel la realitzin com a mínim tres anotadors.
d. S'ensinistra els anotadors en l'ús de l'eina d'anotació i se'ls forma en la temàtica i en els criteris continguts en la guia d'anotació.
e. S'anota el corpus de la mostra i s'identifiquen els acords i els desacords que s'han produït al llarg de l'anotació en paral·lel.
f. Es discuteixen les discrepàncies i s'identifica quines han estat les causes, les quals poden ser:
    i. Una mala interpretació de la guia.
    ii. Especificacions poc clares a la guia.
    iii. Distracció de l'anotador.
g. Es revisa la guia i s'inicia una altra vegada el procés d'anotació en paral·lel.

Aquest procés s'aplica tantes vegades com calgui fins que l'acord entre anotadors és suficientment alt i així s'obté el *Gold Standard*, el corpus definitiu, que servirà de model de qualitat i es podrà utilitzar per entrenar i avaluar algorismes d'Aprenentatge Automàtic.

3. Objectius

En aquest treball ens centrem en el tractament del focus de la negació en documents del domini mèdic per a l'ensinistrament de sistemes de detecció de la negació basats en l'Aprenentatge Automàtic. Volem contribuir en l'estudi del focus de la negació i crear un nou recurs lingüístic, el corpus *ClUB-21* i la guia d'anotació corresponent, que ha de ser útil tant per a estudis teòrics com per a sistemes automàtics entrenats en la identificació del focus de la negació. En concret hem definit les directrius i criteris per anotar la negació i el focus de la negació en el corpus *ClUB-21*.

La nostra motivació sorgeix perquè en l'àrea de l'extracció d'informació l'expressió de la negació encara resulta un aspecte problemàtic i és un tema poc tractat, tot i que la seva identificació és important per comprendre correctament els textos. El focus de la negació indica què és allò negat i, per tant, és fonamental en l'extracció d'informació. A més, com treballem amb textos d'un domini específic, cal també un tractament especialitzat d'aquest fenomen lingüístic.

En aplicar els criteris proposats per Taulé et al. (2020) per a la identificació del focus de la negació hem observat que hi havia més d'un element susceptible a ser interpretat com a focus. Normalment, un d'aquests elements és una expressió temporal i això ens ha portat a tractar també la temporalitat. Per tant, hem estès la nostra anàlisi a les expressions temporals en la mesura en què aquestes interfereixen i modifiquen les expressions negades i atès que afecten la detecció del focus de la negació.

Un dels nostres objectius és donar solució als problemes d'ambigüitat a l'hora de detectar el focus de la negació quan apareixen expressions temporals.

L'objectiu final d'aquest treball és l'elaboració d'una guia anotació del focus de la negació on s'inclouen solucions als problemes que han aparegut.

S'inclou també una tipologia d'expressions temporals que apareixen en el domini mèdic, segons si aquestes fan referència a un període de temps, a un moment del calendari, a una iteració, etc.

4. Estat de l'art

En aquest apartat presentem breument els diferents corpus del domini mèdic anotats amb informació sobre la negació i la temporalitat en espanyol ja que existeixen recopilacions exhaustives sobre la negació (Jiménez-Zafra et al., 2020) i la temporalitat (Alfattni et al., 2020) en general.

Els corpus que presentem han estat desenvolupats per a l'ensinistrament de sistemes de detecció de la negació i de la temporalitat basats en l'Aprenentatge Automàtic.

4.1. Estat de l'art de la negació

Tot i que la detecció de la negació és fonamental per aconseguir uns bons resultats en els sistemes d'extracció d'informació, és un tema que no s'ha abordat fins fa relativament poc temps. Els primers sistemes orientats a la detecció de la negació en l'àmbit mèdic són dels anys 2000. Inicialment aquests sistemes estaven enfocats a processar registres clínics (Chapman et al., 2001; Mutalik et al., 2001). Des de llavors han aparegut diferents sistemes de detecció de la negació en diverses llengües, tot i que la majoria són en anglès.

Atès que la negació s'expressa de maneres totalment diferents segons de quina llengua es tracta, resulta imprescindible la creació de recursos lingüístics específics per a cada llengua, ja que no es poden aplicar de manera general. Malgrat la importància d'aquest fenomen, especialment per a l'extracció d'informació en el domini mèdic, són pocs els corpus anotats amb negació per a altres llengües que no siguin l'anglès.

A continuació presentem els corpus de l'espanyol anotats amb informació sobre la negació en el domini mèdic amb l'objectiu d'avaluar quins aspectes considerem més rellevants i proposar un sistema d'anotació que superi les mancances dels sistemes actuals i així millorar els sistemes d'anotació automàtica basats en l'Aprenentatge Automàtic.

a) El corpus *IxaMed-GS*

Es tracta d'un dels primers sistemes que van anotar la negació en documents mèdics escrits en espanyol (Oronz et al., 2015). L'*IxaMed-GS* és un corpus compost per 75 historials mèdics electrònics de l'Hospital Galdakao-Usansolo de Biscaia. Va ser anotat per dos experts en farmacologia i en farmacovigilància amb l'objectiu d'identificar entitats i esdeveniments continguts en informes mèdics.

Durant el procés d'anotació van considerar la necessitat de marcar tant la negació com la incertesa o especulació, atès que, si una entitat estava negada o es tractava només d'una conjectura, no es podia considerar que formés part de les entitats a tenir en compte en l'extracció d'informació o, en tot cas, s'havia de tractat de manera diferent a les entitats afirmades.

En aquest corpus es van anotar quatre tipus d'entitats: les malalties, les al·lèrgies, els medicament i els procediments clínics. Per a l'anotació de les malalties i les al·lèrgies es distingia entre entitats negades, entitats especulades i entitats no negades ni especulades.

Es van realitzar proves d'acord entre els anotadors i van obtenir un resultat d'un 93,53% d'acord en l'anotació de les entitats i un 82,86% en l'anotació dels esdeveniments. Tenint en compte la dificultat de la tasca, es consideren uns resultats acceptables.

b) El corpus *UHU-HUVR*

Cruz Díaz et al. (2017) van anotar el corpus *UHU-HUVR*, format per 604 informes clínics de l'Hospital de la Virgen del Rocío de Sevilla. Aquest és el primer corpus en espanyol en què es van anotar tots els diferents tipus de negació: sintàctica, lèxica i morfològica.

Es van anotar els marcadors de negació, l'abast de la negació i els esdeveniments negats seguint la guia del corpus *Thyme* (Styler IV et al., 2014). En aquest cas l'acord fou de més del 0.94 en els marcadors i més del 0.72 en els esdeveniments negats.

c) El corpus *IULA Spanish Clinical Records* (*IULA-SCRC)*

L'any 2017 es va anotar també el corpus *IULA Spanich Clinical Records* (*IULA-SCRC*) (Marimon et al., 2017), el qual està format per 300 informes clínics d'un dels hospitals més importants de Barcelona, però no s'especifica quin. En aquest cas no es marca la negació morfològica i es van anotar només els marcadors de negació i l'abast per tres lingüistes computacionals assessorats per un metge clínic. L'acord en l'anotació va ser del 0.85 entre els anotadors 1 i 2 i del 0.88 entre els anotadors 2 i 3.

d) Corpus d'urgències (Madrid)

Des del 2017 s'està treballant encara en l'anotació d'un corpus format per 354.677 informes que són notes d'admissió a urgències d'un hospital de Madrid (Campillos Llanos et al., 2017). El seu objectiu és extreure patrons dels marcadors de negació.

e) Corpus d'informes radiològics

Cotik et al. (2017) han anotat un corpus format per 513 informes radiològics. Van anotar, concretament, les entitats clíniques, la negació, la incertesa i les relacions. L'acord fou del 0.89.

f) El corpus *NUBES*

Lima et al. (2020) han anotat el corpus *NUBES* (*Negation and Uncertainty annoations in Biomedical texts in Spanish*) format per 29.682 oracions extretes de registres mèdics. A banda dels marcadors de negació i del seu abast, en aquest corpus s'anoten també els esdeveniments i la polaritat, és a dir, els elements que reforcen la negació. Per a la seva anotació van partir de la guia del corpus *IULA-SCRC*, la qual van ampliar i modificar.

Valoració:

Com es pot observar, doncs, disposem de pocs sistemes de detecció de la negació en el domini mèdic en espanyol i cap en català. Tres d'aquests projectes (Oronz et al., 2015; Cotik et al., 2017; Lima et al., 2020) tracten també l'anotació de la incertesa amb la finalitat de distingir entre entitats o esdeveniments negats, especulats i reals.

Tots aquests corpus tenen anotades la negació sintàctica i la lèxica, tanmateix alguns prescindeixen d'anotar la negació morfològica: és el cas del *IxaMed-GS* i el *IULA-SCRC*. Només al corpus *NUBES* s'anoten tots els elements que conformen la negació, excepte el focus: els marcadors, l'abast, els esdeveniments i els elements que modifiquen la polaritat. La resta solen anotar només els marcadors i l'abast (*UHU-HUVR* i *IULA-SCRC*) o bé les entitats i els esdeveniments negats (*IxaMed-GS* i el corpus d'informes radiològics).

El valor de les proves d'*Inter-Annotator Agreement* a tots els corpus és d'entre el 80% i el 90%. Tanmateix superen el 90% d'acord l'anotació de les entitats al *IxaMed-GS* i dels marcadors de negació al corpus *UHU-HUVR*.

D'altra banda, cap d'aquests corpus tracta el focus de la negació. És per això que hem consultat també treballs que anoten el focus de la negació en anglès i en espanyol encara que no pertanyin al domini mèdic. Són un total de sis corpus:

EN ANGLÈS:

a) *PropBank Focus* (*PB-FOC)*

El primer corpus en què es va anotar el focus de la negació és el *PropBank Focus* (*PB-FOC*) (Blanco i Moldovan, 2011), el qual està format per 3.993 negacions extretes del corpus *PropBank* de l'anglés (Palme et al., 2005). Fou anotat per dos graduats en lingüística computacional i, un cop anotat el 50%, l'acord fou del 70%. Aquest corpus es va utilitzar posteriorment per Morante i Blanco (2012) com a corpus d'entrenament i de prova per a la detecció del focus durant la primera edició d'una conferència sobre semàntica lèxica i computacional (*SEM 2012). D'altra banda, també Anand i Martell (2012) van reanotar 2.304 exemples del *PB-FOC*.

b) *Deep Tutor Negation* (*DT-Neg)*

El *Deep Tutor Negation* (*DT-Neg*) (Banjade i Rus, 2016) és un corpus format per textos extrets de diàlegs en què estudiants de física interactuen amb Sistemes de Tutoria Intel·ligent (ITS), eines d'ensenyança basades en les TIC que determinen la seqüència i la presentació de continguts en base del rendiment dels estudiants. En aquest cas s'anotà només la negació sintàctica i lèxica (i no la morfològica), els marcadors, l'abast i el focus. L'acord fou d'un 89,43% en l'anotació de l'abast i d'un 94,20% en la del focus.

c) *SFU Opinion and Comments Corpus* (*SOCC*)

El corpus de l'anglès *SFU Opinion and Comments Corpus* (*SOCC*) (Kolhatkar et al., 2019) conté 10.339 articles d'opinió i els seus 663.173 comentaris del diari canadenc *The Globe and Mail*, des de gener del 2012 fins al desembre del 2016. El corpus es va crear amb l'objectiu d'estudiar diferents aspectes dels comentaris on-line com la seva relació amb els articles o la relació entre ells. Tot i així es va seleccionar un subconjunt format per 1.043 comentaris per anotar la negació. S'anotà tant el focus de la negació com els marcadors i el seu abast. Per calcular l'acord es van comparar dues anotacions en paral·lel dues vegades: la primera després de realitzar-se el 50% de l'anotació i la segona posteriorment. L'acord fou del 99,0% en l'anotació dels marcadors durant la primera revisió i del 96,4% durant la segona; del 98,0% en l'anotació de l'abast durant la primera revisió i del 94,2% en la segona revisió; del 85,3% en l'anotació del focus de la negació en el primer càlcul i del 75,8% en el segon.

EN ESPANYOL:

d) El corpus *NewsCom*

Atès que l'únic corpus en espanyol on s'anota el focus de la negació és el *NewsCom* (Taulé et al., 2020) serà la nostra principal referència per a l'anotació d'aquest fenomen.

El *NewsCom* està constituït per comentaris que responen a 18 articles diferents de diaris digitals en espanyol. Les autores segueixen a Huddleston i Pullum (2002) en la definició del focus de la negació com l'element més prominent o explícitament negat, la definició més acceptada en el Processament del Llenguatge Natural. A banda del focus, també anoten els marcadors i el seu abast.

Els criteris dels que parteixen per a la identificació del focus són els següents (apartat 5.2.1):

- Criteri del context discursiu.
- Criteri de l'element més oblic.
- Criteri del significat positiu implícit.

En aquest cas, després d'una segona prova, l'acord en l'anotació fou del 97,29%, un valor molt acceptable atesa la dificultat de la tasca.

Valoració:

Podem concloure que, tot i que hi ha diversos corpus en espanyol del domini mèdic que anoten la negació, no tracten, tanmateix, l'anotació del focus de la negació. A més, aquests corpus anoten la negació sintàctica i lèxica, però no la morfològica, com és el cas dels corpus *IxaMed-GS* i *IULA-SCRS*.

Alguns tracten també les entitats especulades per enriquir l'anotació ja que una entitat especulada és una possibilitat o un suggeriment i, per tant, diferent a una entitat negada o real.

La majoria de corpus anoten només els marcadors de negació i l'abast. Només el corpus *NUBES* anota els esdeveniments i la polaritat.

Totes aquestes dades s'il·lustren a la Taula 1.

Els corpus consultats que anoten el focus de la negació no pertanyen al domini mèdic. Aquests corpus, a banda del focus, anoten també els marcadors de negació i l'abast ja que per determinar el focus és important identificar abans l'abast de la negació, atès que el focus és un element dins de l'abast. Com que cap dels corpus que anoten el focus de la negació és del domini mèdic, pel tractament d'aquest element partim de l'anàlisi directa del nostre corpus, tot i que ens basem en els criteris de Taulé et al. (2020).

Finalment, els resultats de les proves d'*Inter-Annotator Agreement*, recollits a la Taula 2, oscil·len entre el 80% i el 90%. En alguns casos, però, és superior: en l'anotació dels marcadors de negació i de l'abast al *SOCC*, dels marcadors al *UHU-HUVR*, del focus de la negació al *DT-Neg* i l'acord en general del *NewsCom* és del 97,29%. Tanmateix, en altres casos l'acord és inferior al 80%: l'acord en l'anotació d'esdeveniments al *UHU-HUVR* és del 72,0%, l'acord general al *PF-FOC* és del 70,0% i l'acord en l'anotació del focus de la negació al *SOCC* és del 75,8%.

Taula 1. Sistemes de la **detecció de la negació i del focus de la negació**:

REFERÈNCIA	CORPUS	INCER-TESA	NEGA-CIÓ SINTÀ-CTICA	NEGA-CIÓ LÈXICA	NEGA-CIÓ MORFO-LÒGICA	MARCA-DORS	ABAST	ESDEVE-NIMENTS	POLARI-TAT	FOCUS	IDIOMA
Oronz et al., 2015	*IxaMed-GS*	X	X	X				X			Espanyol
Cruz Díaz et al., 2017	*UHU-HUVR*		X	X	X	X	X	X			Espanyol
Marimon et al., 2017	*IULA-SCRC*		X	X		X	X				Espanyol
Campillos Llanos et al., 2017			X	X	X	X					Espanyol
Cotik et al., 2017		X	X	X	X			X			Espanyol
Blanco i Moldovan, 2011	*PB-FOC*		X	X	X	X	X	X		X	Anglès
Morante i Blanco, 2012	*PB-FOC*		X	X	X	X	X	X		X	Anglès
Anand i Martell, 2012	*PB-FOC*		X	X	X	X	X	X		X	Anglès
Banjade i Rus, 2016	*DT-Neg*		X	X		X	X			X	Anglès
Kolhatkar et al., 2019	*SOCC*		X	X		X	X			X	Anglès
Taulé et al., 2020	*NewsCom*		X	X		X	X			X	Espanyol
Lima et al., 2020	*NUBES*	X	X	X	X	X	X	X	X		Espanyol

Font: elaboració pròpia

Taula 2. **Valor del Inter-Annotator Agreement**:

REFERÈNCIA	CORPUS	MARCADORS	ABAST	ENTITATS	ESDEVENIMENTS	FOCUS	EN GENERAL
Oronz et al., 2015	*IxaMed-GS*			93,53%	82,86%		
Cruz Díaz et al., 2017	*UHU-HUVR*	0.94			0.72		
Marimon et al., 2017	*IULA-SCRC*						0.85-0.88
Campillos Llanos et al., 2017							
Cotik et al., 2017							0.89
Blanco i Moldovan, 2011	*PB-FOC*						70%
Morante i Blanco, 2012	*PB-FOC*						
Anand i Martell, 2012	*PB-FOC*						
Banjade i Rus, 2016	*DT-Neg*		89,43%			94,20%	
Kolhatkar et al., 2019	*SOCC*	96,4%	98,0%			75,8%	
Taulé et al., 2020	*NewsCom*						97,29%
Salvador et al., 2020	*NUBES*						

Font: elaboració pròpia

4.2. Estat de l'art: expressions temporals

En els últims anys s'han adreçat grans esforços per identificar automàticament informació clínica clau com són els esdeveniments clínics. Cal, però, situar aquests esdeveniments en un context temporal per entendre l'ordre cronològic dels procediments clínics i així facilitar els diagnòstics o els tractaments, la presa de decisions, controlar els costos, millorar la investigació mèdica i promoure la qualitat del servei. Per això són necessaris sistemes de detecció d'estructures temporals, d'esdeveniments i de les relacions entre aquests elements.

Els principals sistemes d'anotació de la temporalitat parteixen del TimeML, un esquema d'anotació creat amb l'objectiu de definir un llenguatge de marcatge estàndard per als esdeveniments temporals, el temps en què es troben ancorats els esdeveniments i el seu ordre temporal. TimeML s'inicià el 2002 durant els tallers TERQAS (Time and Event Recognition for Question Answering) organitzats pel professor James Pustejovsky amb l'objectiu de millorar els sistemes de pregunta-resposta en llenguatge natural. TimeML es centra en el tractament de preguntes basades en el temps dels esdeveniments i de les entitats. Es disposa de tres versions: TimeML 1.1, TimeML 1.2 i TimeML 1.2.1.

Hem consultat l'última versió, el **TimeML 1.2.1** (Saurí et al., 2006), on s'anoten:

- Els **esdeveniments** amb l'etiqueta <EVENT>, dels quals s'especifiquen diferents atributs com el tipus o classe d'esdeveniment.
- Les **expressions temporals** amb l'etiqueta <TIMEX3>, de les quals s'indica el tipus segons si es tracta: a) d'expressions que fan referència a un moment del dia; b) a un temps de calendari; c) a una extensió en el temps; d) a la freqüència amb què succeeix l'esdeveniment.
- **Elements textuals que fan explícita la relació entre dues entitats**, és a dir: a) entre una expressió temporal amb una altra; b) entre una expressió temporal i un esdeveniment; c) entre dos esdeveniments. Aquests elements poden ser preposicions, conjuncions o caràcters especials i s'anoten amb l'etiqueta <SIGNAL>.
- Per a marcar els **verbs** quan fan referència a dos esdeveniments com, per exemple, "Vaig al cinema els dilluns i els dimarts" s'utilitza l'etiqueta <MAKEINSTANCE>.
- Tres tipus diferents de **relacions entre les entitats**:
    a. La relació temporal amb l'etiqueta <TLINK>.
    b. La relació de subordinació amb l'etiqueta <SLINK>.
    c. La relació aspectual amb <ALINK>.

Taula 3. **TimeML 1.2.1**:

ETIQUETES	ATRIBUT	VALORS
EVENT	eid (Event ID number)	
	Class	REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, STATE, OCCURRENCE
	Stem	
TIMEX3	tid (Timex ID number)	
	Type	DATE, TIME, DURATION, SET
	Value	
	Mod	
	temporalFunction	
	anchorTimeID	
	valueFromFunction	
	functionInDocument	
	beginPoint/ endPoint	
	quant/freq	
SIGNAL	sid (Signal ID number)	
MAKEINSTANCE	eiid (Event Instance ID number)	
	eventID (Event ID number)	
	Tense	PAST, PRESENT, FUTURE, NONE, INFINITIVE, PRESPART, PASTPART
	Aspect	PROGRESSIVE, PERFECTIVE, PERFECTIVE_PROGRESSIVE, NONE
	pos (part-of-speech)	ADJECTIVE, NOUN, VERB, PREPOSITION, OTHER
	Polarity	
TLINK	relType	BEFORE, AFTER, INCLUDES, IS_INCLUDED, DURING, SIMULTANEOUS, IMMEDIATELY AFTER, IMMEDIATELY BEFORE, INDENTITY, BEGINS, ENDS, BEGUN_BY, ENDED_BY
SLINK	relType	MODAL, EVIDENTIAL, NEG_EVIDENTIAL, FACTIVE, COUNTER_FACTIVE, CONDITIONAL
ALINK	relType	INITIATES, CULMINATES, TERMINATES, CONTINUES, REINITIATES

Font: elaboració pròpia

En la taula 3 s'il·lustren les entitats que anota l'última versió del TimeML. En la columna 'ETIQUETES' s'exposen les etiquetes per anotar les diferents entitats, en la columna 'ATRIBUT' els diferents atributs possibles per a cada etiqueta i en la columna 'VALORS' els valors possibles que pot tenir cada atribut.

Per tal de demostrar la viabilitat d'aquest esquema d'anotació, s'anotà **el corpus *TimeBank*** (Pustejovsky et al., 2003). Aquest corpus està format per 300 articles de notícies extretes de diferents fonts com són l'*ABC News* o el *New York Times*. El corpus s'anotà en dues fases. En la primera cinc persones van anotar 210 documents, que suposaven un 70% del corpus. Els anotadors havien participat en el desenvolupament del TimeML; en la segona fase 45 estudiants de Ciències de la Computació van anotar el 30% restant. Per avaluar la utilitat del TimeML per a aplicacions de pregunta-resposta, es creà un corpus format per 50 preguntes, el qual es va anotar també d'acord amb aquest esquema.

Els principals sistemes de detecció de la temporalitat en el domini mèdic parteixen del TimeML, però introdueixen algunes modificacions respecte de l'esquema original. Per exemple, es redueixen les entitats anotades i els seus atributs en base als interessos de cada estudi i projecte. A continuació presentem els projectes consultats.

a) El corpus *i2b2*

L'*i2b2* (Sun et al., 2013a,b) és el primer corpus mèdic que va anotar la temporalitat, el qual segueix una guia d'anotació més senzilla i simplificada que l'esquema del TimeML i es centra en els esdeveniments, les estructures i les relacions temporals. Aquest corpus s'ha constituït com el corpus de referència en l'anotació de la temporalitat en informes mèdics i, per tant, serà la nostra principal referència per l'estudi de la temporalitat. El corpus està format per 310 informes del Partners Healthcare i del Bath Israel Deaconess Medical Center, dos hospitals de Boston. S'anoten tres tipus d'entitats: els esdeveniments clínics, les estructures temporals i les relacions temporals.

Pel que fa als esdeveniments clínics s'especifica:

- **El tipus**, on es distingeixen: malalties, com un càncer o una diabetis; proves clíniques, com una anàlisi de sang; tractaments, com una operació; departaments clínics, com la UCI; evidències, com el verb 'queixar-se'; esdeveniments que li succeeixen al pacient, com un ingrés o una admissió.
- **La polaritat**, és a dir, si l'esdeveniment està negat o no.
- **La modalitat**, on es distingeix si l'esdeveniment és: real, hipotètic, proposat o condicional.

Pel que fa a les estructures temporals s'indica:

- **El tipus** seguint la mateixa classificació que el TimeML.
- **El valor estàndard** segons la norma ISO 8601.
- **El tipus de valor temporal** segons si aquest és exacte o no.

Pel que fa a les relacions temporals s'indica:

- **El tipus**, on s'indica si les dues entitats són simultànies, una és anterior o posterior que l'altre, etc.

En l'anotació de l'*i2b2* és va aconseguir un acord del 0,83 en l'anotació dels esdeveniments, un 0,73 en la de les expressions temporals i un 0, 79 en la de les relacions temporals.

b) El corpus *HPI*

Galescu i Blaylock (2012) van desenvolupar el corpus *HPI*, format per notes clíniques. Per a la seva anotació van seguir la guia de l'*i2b2* amb alguna petita modificació. Concretament canvien els possibles valors que pot prendre la relació temporal.

c) El corpus *THYME*

El *THYME* (*Temporal Histories of Your Medical Events*) (Styler IV et al., 2014) és un corpus format per notes clíniques. Segueix també l'esquema del TimeML i, per tant, coincideix en diversos aspectes amb l'anotació de l'*i2b2*, però introdueix alguns canvis. En aquest corpus s'anoten:

- **Els esdeveniments clínics**. La principal diferència en l'anotació dels esdeveniments respecte de l'*i2b2* és que el *THYME* inclou dins d'aquesta entitat l'atribut per especificar la relació temporal de l'esdeveniment amb una altra entitat.
- **Les estructures temporals**, de les quals també se n'indica el tipus.
- **Les relacions temporals**.
- **Les relacions aspectuals**, de les quals també se n'indica el tipus.

d) El corpus *TempEval*

Finalment, Bethard et al. (2017) van anotar el corpus *TempEval*, el qual està format per notes clíniques i per informes d'anatomia patològica de la Mayo Clinic, un grup de recerca i pràctica mèdica. Segueixen l'anotació del *THYME*, però no anoten les relacions aspectuals, sinó que dels tres tipus de relacions que apareixen al TimeML només anoten les relacions temporals, de la mateixa manera que succeeix al *i2b2*.

Taula 4. **Corpus del domini mèdic que anoten la temporalitat**:

REFERÈNCIA	ETIQUETA	ATRIBUT	VALORS
Sun et al., 2013a,b (*i2b2*)	EVENT	Type	PROBLEM, TEST, TREATMENT, CLINICAL_DEPT, EVIDENTIAL, OCCURRENCE
		Polarity	pos, neg
		Modality	Factual, hypothetical, hedged, conditional
	TIMEX3	Type	DATE, TIME, DURATION, FREQUENCY
		Value	
		Modiefier	MORE, LESS, APROX, START, END MIDDLE, N/A
	TLINK	Type	BEFORE, AFTER, SIMULTANEOUS, OVERLAP, BEGUN_BY, ENDED_BY, DURING, BEFORE_OVERLAP
Galescu i Blylock, 2012	EVENT		
	TIMEX3	Type	DATE, TIME, DURATION, SET
		Value	
	TLINK	relType	BEFORE, AFTER, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER, VAGUE
Styler IV et al., 2014 (*THYME*)	EVENT	DocTimeRel	BEFORE, OVERLAP, AFTER, BEFORE-OVERLAP
		Type	N/A, ASPECTUAL, EVIDENTIAL
		Polarity	POS, NEG
		Degree	N/A, MOST, LITTLE
		Contextual modality	ACTUAL, HEDGED, HYPOTHETICAL, GENERIC
		Contextual aspect	N/A, NOVEL, INTERMITTENT
		Permanence	FINITE, PERMANENT, UNDETERMINED
	TIMEX3	Class	DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP, SET
	TLINK	Type	BEFORE, OVERLAP, BEGINS_ON, ENDS_ON
	ALINK	Type	CONTINUES, INITIATES, REINITIATES, TERMINATES

Bethard et al., 2017 (*Clinical TempEval*)	EVENT	DocTimeRel	BEFORE, OVERLAP, AFTER, BEFORE-OVERLAP
		Type	N/A, ASPECTUAL, EVIDENTIAL
		Polarity	POS, NEG
		Degree	N/A, MOST, LITTLE
		Modality	ACTUAL, HEDGED, HYPOTHETICAL, GENERIC
		Contextual aspect	N/A, NOVEL, INTERMITTENT
	TIMEX3	Class	DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP, SET
	TLINK	Type	

Font: elaboració pròpia

Valoració:

Podem concloure que tots aquests corpus parteixen del TimeML, tots són en anglès i tots coincideixen en l'anotació de:

- **Els esdeveniments**, marcant de quin tipus són, és a dir, si es tracta d'un tractament, d'una prova, d'un departament clínic, etc.
- **Les estructures temporals** amb l'etiqueta <TIMEX3>, de les quals també s'indica el tipus segons si es tracta d'expressions que fan referència a un moment del dia, a un temps de calendari, a una extensió en el temps o a la freqüència amb què succeeix l'esdeveniment.
- **Les relacions temporals** amb l'etiqueta <TLINK>, indicant-ne també el tipus de relació.

Només el *THYME* anota, a més, les relacions aspectuals (ALINK), un altre dels tres tipus de relacions que apareixen al TimeML. D'altra banda, en la majoria de corpus també s'indiquen altres característiques dels esdeveniments com són la polaritat o la modalitat, és a dir, si es tracta d'un esdeveniment condicional, hipotètic, factual, etc., la qual cosa està relacionada amb la negació i l'especulació. Un esdeveniment hipotètic o condicional és un esdeveniment especulat i un esdeveniment absent és un esdeveniment negat.

5. Metodologia

En aquest apartat exposem la metodologia que hem seguit en el nostre treball. Es presenten les característiques dels informes mèdics que composen el corpus *ClUB-21* ja que és la base sobre la qual hem desenvolupat la nostra proposta d'anotació de la negació i de la temporalitat en el domini mèdic. Analitzem la negació i la temporalitat i la problemàtica que plantegen en el domini mèdic.

5.1. El corpus ClUB-21

El corpus *ClUB-21* consisteix en un conjunt d'informes mèdics que ens ha cedit l'Hospital Clínic de Barcelona, un hospital universitari i públic situat a la part esquerra de l'Eixample i guardonat per IASIST com un dels 20 millors hospitals d'Espanya. El corpus conté 24 informes de curs clínic de diversos pacients i els seus corresponents 24 informes d'alta. Com que es tracta de documents privats i confidencials, d'acord amb la llei 41/2002, ens han estat cedits un cop han estat anonimitzats.

El nostre corpus, per tant, està conformat per dos tipus diferents d'informes: els informes de curs clínic i els informes d'alta.

Els informes de curs clínic són els informes que es van redactant i ampliant a partir del moment en què el pacient ingressa a l'hospital. Solen ser més llargs i amb una estructura més variable que els informes d'alta. A més, en la seva redacció hi intervenen diferents facultatius. L'estructura més habitual és la següent (Imatge 1):

- S'exposen les dades personals del pacient. En aquest apartat es recopilen dades i informació sobre el pacient i en el nostre corpus per motius de confidencialitat s'han eliminat algunes d'aquestes dades (línies 3, 4 i 5).
- S'explica el motiu de l'ingrés. Aquest apartat serveix per explicar les raons per les quals el pacient acudeix al centre mèdic (línia 8).
- S'exposen els antecedents personals. En aquest apartat s'exposa informació d'interès com poden ser malalties prèvies, al·lèrgies o hospitalitzacions prèvies del pacient (en l'exemple de la Imatge 1 no se'n donen).
- Es fan valoracions. Aquestes valoracions es van realitzant al llarg de tot l'ingrés i estan organitzades en seccions. En cada valoració s'especifica el dia i l'hora i s'anoten diferents tipus de dades: proves que se li realitzen al pacient i els resultats, medicació que se li administra, diagnòstics, etc. (línies 19-27).

Sovint, però, aquests apartats es repeteixen més d'una vegada al llarg del document. A més, freqüentment ens trobem amb l'elisió d'algun apartat i un ordre variable.

Els informes d'alta, d'altra banda, resumeixen l'assistència dispensada al pacient i informen sobre el diagnòstic, el tractament i les recomanacions que ha de seguir. Solen ser més breus i

segueixen una estructura més clara i fixa. Reaprofiten la informació dels informes de curs clínic, però resumida i estructurada. L'estructura dels informes d'alta sol ser la següent (Imatge 2):

- Es recopilen les dades personals del pacient (línies 3, 4 i 5).
- S'explica el diagnòstic, és a dir, es determina quina malaltia o condició pateix el pacient (línies 11-14).
- S'exposa el motiu de la consulta i els antecedents personals (línies 17-24).
- Es narra l'evolució clínica (línies 31-39).
- Es descriuen l'exploració física i les proves realitzades, així com els resultats (línia 26).
- S'explica quin tractament ha de seguir el pacient (línies 41-44).

Imatge 1. Exemple d'un **informe de curs clínic** del corpus *ClUB-21*

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <CursClinic>
3    <sex>2</sex>
4    <nhc>0000020943</nhc>
5    <age>66 años</age>
6    <NotaClinica>
7      <nota lang='es'>(VER NOTA DE INGRESO)
8  1.- NEUMONIA BASAL IZQUIERDA. PROBABLE INFECCIÓN POR COVID 2.- HIPOCELCEMIA LEVE
9  ----------
10 Buen estado general. Eupneica, afebril y saturación basal 95%. Rx con infiltrado en base izquierda.
11 Analítica con elevación RFA (PCR 24), colestasis disociada, hipocalcemia 7.6, neutrofilia y linfopenia. QTc 420ms.
12 Se inicia triple terapia + cefriazona + calcio ev. Analítica para lunes.</nota>
13     <date>28.03.2020</date>
14     <time>21:43</time>
15     <prof>MED</prof>
16     <serv>INF</serv>
17   </NotaClinica>
18   <NotaClinica>
19     <nota lang='es'>paciente consciente y orientada. Normotensa y febricular por la noche.
20     Realiza 3 depos líquidas y refiere dolor abdominal, que después cede.
21 Sat 96% en basal.</nota>
22     <date>29.03.2020</date>
23     <time>07:21</time>
24     <prof>ENF</prof>
25     <serv>INF</serv>
26   </NotaClinica>
27 </CursClinic>
```

Font: elaboració pròpia

Imatge 2. Exemple d'un **informe d'alta** del corpus *ClUB-21*

```
<?xml version="1.0" encoding="UTF-8"?>
<iaHosp>
<ID>1006752517</ID>
<SEX>2</SEX>
<AGE>60a</AGE>
<SPECIALITY>HDM</SPECIALITY>
<DATE_INI>2020/04/08</DATE_INI>
<TIME_INI>16:39:17</TIME_INI>
<DATE_FIN>2020/04/21</DATE_FIN>
<TIME_FIN>11:51:13</TIME_FIN>
<DIAGS>
<DIAG lang='es'>COVID-19</DIAG>
<DIAG lang='es'>Infección por SARS-COVID19</DIAG>
</DIAGS>
<PROCS>
</PROCS>
<MOTIVO_DE_CONSULTA lang='es'>
Infección por SARS-COVID19
</MOTIVO_DE_CONSULTA>
<ANTECEDENTES lang='es'>
Antecedentes personales: niega
Alergias: Penicilina
No toma tratamiento de forma habitual.
</ANTECEDENTES>
<PROCESO_ACTUAL lang='es'>
Ingresa en Hospitalización Domiciliaria derivado de Salut Laboral por Frotis positivo para SARS-COVID-19.
</PROCESO_ACTUAL>
<EXPLORACION_CLINICA lang='es'>
.
</EXPLORACION_CLINICA>
<EVOLS>
<EVOL>
<DATE></DATE>
<TIME>000000</TIME>
<TEXT lang='es'>
Paciente con infección asintomática por SARS-COVD-19.
Se pauta aislamiento domiciliario durante mínimo 14 días con buena evolución clínica, se da de alta de la Hospitalización Domiciliaria.
</TEXT>
</EVOL>
</EVOLS>
<PLAN_TERAPEUTICO lang='es'>
Utilizar mascarilla quirúrgica en puesto de trabajo y en domicilio durante 1 semana.
Ante cualquier sintomatología contactar con Salud Laboral.
</PLAN_TERAPEUTICO>
<SEGUIMIENTO lang='es'>
</SEGUIMIENTO>
</iaHosp>
```

Font: elaboració pròpia

Aquests documents ens han estat cedits en format TXT i anotats amb XML. Els informes d'alta tenen metadades per assenyalar els diferents apartats i els de curs clínic només marquen les diferents anotacions que es van escrivint al llarg de l'ingrés. Ambdós, però, marquen també algunes dades específiques com l'edat o el sexe del pacient. Pel que fa a la temporalitat, els informes de curs clínic anoten amb les etiquetes <date> i <time> el dia i l'hora en què es redacta una nota clínica. Els informes d'alta anoten amb les etiquetes <DATE_INI>, <TIME_INI>, <DATE_FIN> i <TIME_FIN> el dia i l'hora de l'ingrés i el dia i l'hora de l'alta. No s'anota cap altra dada temporal. Aquesta informació marcada en XML facilita l'extracció posterior de les estructures temporals anotades.

Es tracta, per tant, de documents semiestructurats i no estructurats que precisen de tècniques d'extracció d'informació per a l'obtenció automàtica d'informació estructurada.

5.2. La negació en el domini mèdic

La negació és un fenomen complex que s'ha estudiat des de diferents perspectives: la filosofia, la lingüística i, fins i tot, la psicologia. Des d'un punt de vista lingüístic ens trobem amb què es pot expressar mitjançant recursos sintàctics, morfològics i lèxics. En català i castellà, la forma més habitual i freqüent d'expressar la negació és mitjançant un marcador. Els marcadors de negació normalment modifiquen el valor de veritat d'una expressió, però poden tenir altres valors semàntics com, per exemple, marcar l'absència d'un esdeveniment o d'una entitat. Per exemple:

(23)

No s'administra bisoprolol per precaució

(24)

Abdomen: blando, depresible, **sin** dolor

En l'exemple (23) s'observa com l'adverbi 'no' modifica el contingut de l'oració per indicar que no es produeix un determinat esdeveniment, mentre que en l'exemple (24) la preposició 'sin' indica l'absència de l'entitat 'dolor'.

Tot i que es tracta d'un fenomen present a totes les llengües, la negació s'expressa de maneres diferents a cada una d'elles, motiu pel qual requereix un tractament específic per a cada una. Per exemple, les oracions "No té cap esperança" en català o "No tiene ninguna esperanza" en castellà són equivalents a "He/she has no hope", però en català i castellà s'utilitzen dos elements negatius ('no' i 'cap' o 'ninguna') mentre que en anglès només un ('no').

En segons quins contextos i també en segons quins dominis temàtics, la negació pot contenir un significat positiu implícit i, per tant, aporta també informació sobre l'existència de determinats fets o fenòmens. Per exemple:

(25)

No tuvo contacto con otros enfermos

L'oració (25) implica que va tenir contacte amb altres persones, però no amb malalts. Trobem, doncs, un significat afirmatiu, és a dir, un coneixement vertader implícit que és que "no estava aïllat i havia tingut contactes". Només es nega "con otros enfermos". En aquest cas, l'element negat és el que s'anomena focus de la negació, el tractament del qual és un dels objectius d'aquest treball.

En l'àmbit del Processament del Llenguatge Natural i, en concret, de l'extracció d'informació, el focus de la negació ha estat poc tractat fins ara: normalment s'han tractat només el marcador de negació i l'abast.

En els exemples següents, tots ells procedents del nostre corpus, hem marcat tots els elements que poden intervenir en les expressions que contenen negació: el marcador en **negreta**, l'abast entre [claudàtors], l'esdeveniment en *cursiva*, els modificadors de polaritat entre (parèntesis) i el focus subratllat:

(26)

a. **No** [*tomaba* (ninguna) medicación]
b. **No** [se le *hizo* (ninguna) prueba inicial de SARS-CoV-2]
c. **No** [se *hizo* la prueba de SARS-CoV-2 a (ninguno) de los contactos]
d. **No** [*reveló* (ninguna) arteriopatía coronaria obstructiva]

És important distingir entre 'negació' i el significat positiu o negatiu d'una expressió, especialment en el domini mèdic. Moltes vegades en una història clínica que un fet estigui negat pot implicar un valor positiu per a l'estat general del pacient:

(27)

PCR coronavirus: **Negativo**

En el nostre treball no entrarem en aquest tipus d'implicacions. Ens limitarem a detectar les expressions que indiquen negació i a anotar els seus components, independentment del valor positiu o negatiu que puguin tenir per a l'estat del pacient.

A més, les estructures negatives poden referir-se a una àmplia gama de situacions o processos relacionats amb el pacient, com són ara la finalització d'un procés o d'un estat (28a), l'absència d'un determinat símptoma o d'una malaltia (28b), la no realització d'un esdeveniment (28c), etc.

(28)

a. **Finaliza** el tratamiento
b. **Ausencia de** marcadores de riesgo
c. **Ni** come **ni** bebe

L'àmplia gama de possibles valors de les estructures que indiquen negació i l'impacte semàntic que provoca en les expressions que la contenen posa en evidència la importància de la detecció d'aquest fenomen, especialment en el domini mèdic on està en joc la vida de persones.

5.2.1. Tipus d'estructures negatives en el domini mèdic

A continuació presentem les diferents estructures negatives que apareixen en els informes mèdics.

a) Negació sintàctica

Normalment l'expressió de la negació sintàctica es realitza mitjançant una paraula funcional. En el domini mèdic, les més freqüents són l'adverbi 'no' i la preposició 'sin'/'sense' segons la llengua.

Trobem les estructures següents:

- L'adverbi 'no' seguit d'un substantiu. Es dona especialment en estructures en què s'ha elidit el verb, molt freqüents al domini mèdic, com ja s'ha vist.

(29)

a. **No** hábitos tóxicos
b. **No** ruidos agregados
c. **No** dèficit motor

- L'adverbi 'no' seguit d'un adjectiu o d'un participi.

(30)

a. Vía biliar **no** dilatada
b. Tos **no** productiva
c. Un gran patrón **no** isquémico

- L'adverbi 'no' seguit d'un verb o d'una perífrasi verbal.

(31)

a. **No** hi ha dolor
b. **No** presenta dificultad respiratoria
c. **No** se palpan masas

- La preposició 'sense' en català o 'sin' en castellà seguida d'un sintagma nominal o d'un infinitiu.

(32)

a. Ambas ilíacas externas mestran segmentos **sin** calcificaciones
b. **Sin** incidentes durante el procedimento
c. La familia de había autoconfinado a causa de la pandemia de COVID-19 durante la semana anterior, **sin** salir para ir a la escuela ni al trabajo

b) Negació lèxica

La negació lèxica és la que s'expressa mitjançant paraules que pertanyen a una de les categories gramaticals obertes: noms, verbs, adjectius i/o participis. La forma lèxica conté implícita el contingut de negació.

La casuística que presenta és la següent:

- El marcador lèxic és un substantiu seguit d'una preposició. Normalment expressa un canvi d'estat o la finalització d'un procés.

(33)

a. Paciente con **negación a** requerimiento cercando de TRS
b. **Ausencia de** síntomas respiratorios
c. **Desaparición** completa **de** los síntomas

- El marcador és un verb que expressa un canvi d'estat o la finalització d'un procés o un *verba dicendi* com 'negar'.

(34)

a. **Niega** consumo de hábitos tóxicos

b. **Es finalitza** administració de dos concentrats
c. Tras 48 horas, los síntomas **remitieron**

- El marcador és un adjectiu que expressa un canvi d'estat o la finalització d'un procés o, en el cas de resultat analítics, l'adjectiu 'negatiu' o 'negativo' que indica el valor del resultat d'unes proves.

(35)
a. Coprocultivo **negativo** y toxina clostridium **negativa**
b. Serologías para HD **negatives**
c. Síntomas **ausentes**

c) Negació morfològica

La negació morfològica és aquella que s'expressa mitjançant afixos generalment d'un nom, un verb o d'un adjectiu.

(36)
a. **A**febril mantiene saturación 95%
b. Frotis COVID 19: **In**detectable
c. Estat de la consciència: **In**conscient
d. Paciente C y O, HDM estable con tendencia a la hiperTA **a**sintomática y **a**febril

5.2.2. Problemes específics de la negació en el domini mèdic

En aquest apartat exposem algunes qüestions referents a la negació que presenten característiques peculiars en el domini de coneixement representat pel corpus *ClUB-21*. Aquestes qüestions són específiques del subllenguatge mèdic i no es presenten necessàriament en la llengua estàndard.

- Un aspecte del domini mèdic que cal destacar pel que fa a la negació és la abundància de marcadors de negació lèxics i morfològics. Això s'explica, entre d'altres raons, pel fet que el subllenguatge mèdic presenta una estructura sintàctica molt simplificada, el que fa que predominin les expressions nominals i adjectives.

(37)
a. **A**febril, **ausencia de** síntomas respitarorios.
b. **A**sintomàtico. **Niega** síntomas, por lo que **finalizo** tratamientio

- A més, des d'una perspectiva semàntica, cal remarcar que en els informes mèdics la negació moltes vegades s'ha d'interpretar com l'absència o la supressió d'un determinat símptoma, d'una malaltia o d'un esdeveniment. És menys freqüent l'ús de la negació com a indicador de la no veritat d'un enunciat.

(38)
a. **Descartar** sobreinfección bacteriana

b. **Retiro** esparadrapo
c. La persona de contacte **no** correspon a cuidador

En els exemples (38a) i (38b), els verbs 'descartar' i 'retiro' són marcadors lèxics de negació i marquen l'absència de les entitats 'sobreinfección bacteriana' i 'esparadrapo'. En l'exemple (38c), en canvi, l'adverbi 'no' indica la no veritat de l'enunciat "la persona de contacte (no) correspon a cuidador".

- Destaca l'alta freqüència de termes mèdics que des d'un punt de vista etimològic incorporen algun afix que expressa negació. Aquests termes formen part d'un vocabulari culte que s'ha manllevat del llatí i del grec. Es tracta, però, de paraules en què hi ha una integració lèxica d'aquests afixos, pel que en la llengua actual no es perceben com paraules de negació. Per exemple, la paraula 'dispnea' ve del grec, etimològicament està formada pel prefix negatiu 'δυσ-' i 'πνέω' i significa "que respira amb dificultat" o "que no respira". Quan apareix en els informes mèdics considerem que realment no s'està indicant l'absència de respiració, sinó que senzillament és un símptoma que realment no genera negació. Els següents termes són exemples d'aquest tipus de lèxic: 'afàsia', 'anòsmia', etc. Aquests termes no els hem considerat per a l'anotació.

- En front d'aquests termes lexicalitzats, hi ha una àmplia terminologia, en part compartida amb la llengua comuna, que presenta afixos que indiquen negació. Es tracta de marcadors de negació morfològics. Aquests termes sí que els anotem i en donarem una llista a l'apèndix 9.1. Les següents paraules són exemples de negació morfològica: 'afebril', 'asimptomàtic', 'indeterminat', etc.

- La delimitació entre la terminologia lexicalitzada i aquella que entenem com a marcadors de negació és poc definida i pot ser una font de possibles errors.

(39)
a. Pneumònia **a**típica
b. **In**continència

Els termes (39a) i (39b) són uns exemples on es pot presentar aquesta problemàtica. La paraula 'atípica' per sí mateixa significa que una cosa no és típica. Tanmateix en aquest context cal considerar-la com un sol terme juntament amb 'pneumònia' que assenyala un tipus de pneumònia i no com a negació de 'típica'. La paraula 'incontinència' té el prefix negatiu 'in-', però no existeix una paraula positiva en contraposició, pel que no consideraríem que s'estigui negant cap cosa.

- Una altra característica del domini mèdic és l'ús de paraules com a marcadors de negació que en la llengua comuna no tenen un valor de negació. Per exemple, el verb 'retirar' el considerem com a marcador de negació en oracions com "Se retira la medicación" ja que entenem que s'indica l'absència d'una entitat, en aquest cas de la medicació. En la llengua estàndard, en oracions com "retirar dinero del banco" no indica una absència i, per tant, no genera negació.

Algunes d'aquestes paraules poden expressar diferents significats o matisos pel que pot generar-se confusió. Per exemple:

(40)

a. **Se eliminó** [la arritmia mediante cardioversión]
b. **Se eliminó** [un metro de íleon y de colon derecho necróticos]

Tant en l'exemple (40a) com en el (40b) el marcador de negació és el verb 'eliminar', però mentre que en el (40a) significa "fer desaparèixer alguna malaltia o símptoma" i, per tant, marca l'absència d'una entitat, en el (40b) indica "suprimir mecànicament", pel que marca un procés i no una absència. L'oració (40b), per tant, no la considerem una negació.

- Un tipus d'estructures característiques del subllenguatge mèdic són les de valor-resultat, que en alguns casos són estructures predeterminades. En aquests casos, l'abast i el focus de la negació es trobem avantposats i separats per signes de puntuació del marcador de negació:

(41)

a. [Beu de forma habitual]?: **No**
b. [Té dolor]?: **No**
c. [Alteració de la integritat de la pell a l'ingrés]: **No**

5.2.3. El focus de la negació

Pel que fa al focus de la negació, que és l'objectiu principal del nostre estudi, és la part de l'abast que ha de ser interpretada com específicament negada i la seva anotació permet un tractament de la informació més detallat i fiable. A més, la seva detecció és útil també per la recuperació de les paraules que contribueixen a aportar el significat positiu implícit que pot derivar-se d'una negació.

El focus, però, és un element especialment difícil d'identificar atès que per a la seva detecció no sempre es disposa d'indicadors lingüístics formalment explicitats. El focus pot estar determinat per la semàntica dels mots o per aspectes pragmàtics com les intencions comunicatives o, fins i tot, pot dependre d'informació paralingüística. Considerem, per exemple, la següent oració:

(42)

No realiza seguimiento por Hepatología de HCB

En aquesta frase la negació està marcada mitjançant l'adverbi 'no', el qual modifica 'realiza'. Tanmateix el focus de la negació és 'por Hepatología de HCB' ja que no es nega que es realitzi seguiment, sinó que es faci 'por Hepatología de HCB'. Per tant, la forma totalment explícita d'aquest enunciat seria: "Realiza seguimineto, pero **no** por Hepatología de HCB". No podem aplicar, doncs, criteris sintàctics per a la seva identificació ja que sintàcticament 'no' modifica el verb 'realiza'.

Per a la identificació del focus de la negació partim dels criteris de Taulé et al. (2020) aplicats al corpus *NewsCom*, el primer i únic corpus en espanyol que anota el focus. Aquest parteix de tres criteris d'identificació:

- **Criteri del context discursiu**. Segons aquest criteri, per a la correcta identificació del focus de la negació cal tenir en compte el context del discurs i, per tant, tot el document, no només l'estructura negada.

- **Criteri de l'element més oblic**. Aquest criteri estableix que l'element més oblic de l'oració és el candidat més plausible de ser el focus de la negació ja que s'està fent explícita informació específica que, de no tenir un valor informatiu rellevant, probablement no s'explicitaria. Si hi ha un element oblic (normalment els complements circumstancials i els adjunts oracionals) és perquè aporten un contingut rellevant des del punt de vista informatiu i, per tant, constitueixen el principal candidat a ser el focus de la negació.
  (43)
  a. **Sin** [alteraciones destacables en el perfil hepático y pancreático]
  b. **No** [edemas en MMII]

  En aquests exemples, aplicant el criteri de l'element més oblic, consideraríem el focus de la negació els adjunts de lloc (43a,b).

- **Criteri del significat positiu implícit**. Segons aquest criteri el focus de la negació és l'element que ha de ser interpretat com a fals per a fer certa la negació general i, per tant, una afirmació negada pot comportar un significat implícit positiu.
  (44)
  El padre **no** [había tenido contactos conocidos con individuos enfermos]

  D'acord amb aquest criteri, en l'exemple (44), es podria extreure el coneixement cert que el pare havia estat en contacte amb altres individus, però no 'con individuos enfermos', el qual és l'element que ha de ser interpretat com a fals.

A partir d'aquests criteris pragmàtico-semàntics distingeixen entre el focus explícit o implícit.

El focus de la negació explícit és aquell que té un marcador que l'identifica formalment.

Hi ha diferents recursos lingüístics, generalment sintàctics, pragmàtics o tipogràfics, que permeten destacar un element de l'enunciat, que correspon al focus. Aquest procediment s'anomena focalització i consisteix a emfatitzar de manera explícita un element del discurs. El focus de la negació s'explicita mitjançant alguns d'aquests recursos.

Taulé et al. (2020) presenten diversos procediments que permeten identificar de manera explícita el focus de la negació en espanyol:

- **El desplaçament (del focus)**. El desplaçament d'un element és un mecanisme de focalització que consisteix a moure l'element focalitzat a una posició sintàcticament marcada, normalment al davant de tot de l'oració o bé al final després d'una pausa. En un enunciat que conté una negació, si un constituent es troba en una posició marcada és altament probable que sigui el focus de la negació.

  (45)

  Paciente conciente y orientado que ingresa procedente del H. Clinico por disnea, covid +, portador de [*VMK y VE fuera de vena*], **se retira**[3]

  En l'exemple (45) veiem com el focus es troba en una posició anterior al marcador i separat per una pausa (la coma). El trobem, doncs, en una posició marcada.

- **El subjecte pronominal explícit**. En castellà i en català els subjectes pronominals no es solen expressar ja que la forma verbal indica la persona gramatical. Per tant, si apareixen explícits amb verbs intransitius o en construccions contrastives són el focus de la negació. No hem trobat aquest tipus de construcció al corpus *ClUB-21*.

- **Les construccions contrastives**. Aquest tipus de construccions estan introduïdes generalment per conjuncions adversatives que presenten una alternativa a un enunciat anterior que conté una negació. Aquesta alternativa que contrasta amb l'element anterior negat permet identificar aquest element com el focus de la negació. A més, el significat positiu sol estar explícitament expressat en la segona part de la construcció.

  (46)

  No [se realiza angio-TC] *pero se realiza radiografía*

- **Reforçament de la negació**. El reforçament de la negació consisteix en l'ús de construccions amb més d'un marcador de negació. Generalment trobem un adverbi 'no' i un segon marcador que és un pronom o un adverbi. Aquest segon marcador permet identificar el focus. En el nostre corpus, el *ClUB-21*, trobem construccions d'aquest tipus, especialment construccions en les que el focus està marcat mitjançant un pronom o un determinant indefinit:

  (47)

  a. **No** [reveló *ninguna* arteriopatía coronaria obstructiva]
  b. **No** [se hizo la prueba de SARS-CoV-2 a *ninguno* de los contactos]

- **Marcadors tipogràfics**. Mitjançant recursos tipogràfics com les majúscules, la negreta, la cursiva o el subratllat s'emfatitza un element, el qual és el focus de la negació.

  (48)

  [*FROTIS* control tras 3 semanas] **NEGATIVO**

[3] Ens els exemples (45), (46), (47) i (48) es marquen en negreta els marcadors de negació, entre claudàtors l'abast, subratllat el focus de la negació i en cursiva els recurssos per fer eplícit el focus.

Pel que hem observat, en el domini mèdic són poc freqüents alguns d'aquests mecanismes. Les característiques lingüístiques del subllenguatge mèdic fan que, per exemple, no trobem subjectes pronominals explícits atès que s'intenta que els informes mèdics siguin el més objectius possibles. Tampoc trobem gaires construccions contrastives per la simplicitat sintàctica que els caracteritza i la tendència a la juxtaposició de les diferents oracions. Tot i que sí trobem elements topicalitzats, no és freqüent la topicalització del focus de la negació.

Trobem amb més freqüència elements que reforcen la negació i que marquen polaritat. En aquests casos, per tant, considerem que el focus de la negació és el sintagma que conté l'element que marca el reforçament, que sol ser un pronom o un determinant indefinit, com veiem en l'exemple (47a) i el (47b).

Quan **el focus de la negació** és **implícit** i, per tant, no hi ha cap marcador que faciliti la seva identificació, per al seu reconeixement es tenen en compte criteris pragmàtico-semàntics: el context discursiu, l'element més oblic i el significat positiu implícit. Distingim entre les construccions sense predicat verbal i les construccions amb predicat verbal.

a) Construccions sense predicat verbal

En els informes mèdics, les construccions sense predicat verbal són les construccions formades per: i) l'adverbi 'no' seguit d'un substantiu; ii) l'adverbi 'no' seguit d'un adjectiu o d'un participi; iii) les construccions en què el marcador de negació és un substantiu o un adjectiu seguits d'una preposició; iv) la preposició 'sense' o 'sin' seguida d'un sintagma nominal o d'un infinitiu.

b) Construccions amb predicat verbal

En les construccions amb predicat verbal distingim entre aquelles en què el focus de la negació és un argument i aquelles en què és un adjunt.

En primer lloc, trobem construccions on el verb, tant si és perifràstic o no, només està complementat per arguments i no per adjunts. Els diferents tipus de construccions amb predicat verbal en què el focus de la negació és un argument són les següents:

i) Construccions amb verbs intransitius. En aquests casos el focus pot ser el verb o bé el subjecte.

(49)

El paciente **no** [mejora]

ii) Construccions amb verbs existencials. En aquest tipus d'oracions el focus és el subjecte existencial ja que el verb està buit de significat.

(50)

No [hi ha dolor]

iii) Construccions amb verbs transitius, copulatius o que regeixen preposició. En aquests casos el focus de la negació és el complement directe, l'atribut o el sintagma preposicional.

(51)

a. **No** [refiere algias]
b. La anticoag **no** [és imprescindible]
c. **No** [respon a estímuls verbals]

iv) Construccions amb verbs ditransitius, és a dir, que necessiten un complement directe i un altre complement. En aquestes oracions cal aplicar el criteri de l'argument més oblic.

(52)

No [se aplica este tratamiento a las zonas infectadas]

En segon lloc, quan el predicat verbal està complementat per algun adjunt, l'adjunt és l'element més susceptible de ser el focus de la negació ja que és l'element més oblic. Els adjunts aporten informació no exigida pel predicat i, per tant, la seva explicitació denota que és una informació rellevant.

(53)

a. **No** [s'administra bisoprolol per via oral]
b. El paciente **no** [se encuentra en condiciones para entrar a quirófano]
c. Vostè [beu de forma habitual]?: **No**

Ara bé, en aquelles oracions amb predicat verbal en què una expressió temporal és l'adjunt s'ha detectat que es genera ambigüitat a l'hora d'identificar el focus de la negació. Ho tractarem amb més detall a l'apartat 6.

5.3. Expressions temporals en el domini mèdic

La informació temporal té com a objectiu fonamental situar un esdeveniment en un marc temporal.

La detecció i el tractament de la informació temporal és necessària i aporta coneixement fonamental en l'extracció d'informació, els sistemes de pregunta-resposta o el resum automàtic. Per exemple, per a que un sistema de pregunta-resposta pugui respondre correctament a una pregunta com "quant de temps van durar els símptomes?" cal haver detectat en una oració com "els símptomes van durar tres dies" l'expressió 'tres dies' i identificar-la com a temporal. En aquest exemple s'hauria d'anotar l'esdeveniment 'els símptomes van durar', l'estructura temporal 'tres dies' i la relació temporal entre ambdós.

Com ja s'ha explicat a l'apartat 2.2., per al correcte tractament de la informació temporal, cal anotar tant els esdeveniments i les estructures temporals com les relacions temporals entre esdeveniments i estructures temporals.

Per a la correcta identificació dels esdeveniments i de les estructures temporals cal partir d'una tipologia. En el domini mèdic trobem tipus d'esdeveniments i d'estructures temporals que són particulars d'aquest subllenguatge i que presentem a continuació.

a) Els esdeveniments clínics

Els esdeveniments clínics comprenen tota varietat de successos que tenen lloc en l'àmbit hospitalari i que estan referits a un pacient. Es poden expressar mitjançant verbs conjugats, verbs en forma no personal, substantius, adjectius o sintagmes preposicionals.

D'acord amb Sun et al. (2013a) un esdeveniment clínic és qualsevol cosa rellevant per a la línia temporal clínica. Poden agrupar-se en quatre grups:

- **Conceptes clínics i entitats clíniques**. Es tracta d'expressions referents a una malaltia, a un tractament, a proves mèdiques, a símptomes o a medicaments.
  (54)
  a. **Diverticulitis aguda** (2016), **colecistitis aguda** (2018)[4]
  b. No administramos **tratamiento ACO** de esta noche
  c. **PCR para COVID en frotis nasofaringeo**, positiva
  d. **Fiebre de 38ºC** (posteriormente **febrícula**), **astenia**, **artromialgias** y **tos seca**
  e. No precisa de **primperam**

- **Departaments clínics**.
  (55)
  Procedent d'**urgències**

- **Evidències**. Es refereixen per evidències a aquells esdeveniments que revelen la font de la informació. Es tracta, per tant, de *verba dicendi*.
  (56)
  Refiere molestia en hipogastrio, polaquiuria y disuria desde ayer

- **Ocurrències o successos**. Aquest tipus d'esdeveniments són aquells que li succeeixen al pacient.
  (57)
  a. Pacient que **ingressa** a la a càrrec de ONCO per suboclusió intestinal
  b. Posible **alta** esta tarde

[4] En aquest apartat (5.3.) marquen en negreta els esdeveniments i les entitats clíniques.

b) Les estructures temporals en el domini mèdic

Les estructures temporals són aquelles expressions que donen informació sobre el temps en què succeeix un esdeveniment, la seva durada o la seva freqüència.

En els informes mèdics del nostre corpus, el *ClUB-21*, trobem algunes d'aquestes estructures anotades en XML que funcionen com a metadades. Estan marcades aquelles estructures que fan referència al dia i a l'hora en què es redacta una nota de curts clínic (58) i, en els informes d'alta, s'especifica el dia i l'hora de l'ingrés i de l'alta (59):

(58)

<date>20.08.2020</date>[5]
<time>21:57</time>

(59)

<DATE_INI>2020/08/20</DATE_INI>
<TIME_INI>19:51:44</TIME_INI>
<DATE_FIN>2020/08/22</DATE_FIN>
<TIME_FIN>18:30:11</TIME_FIN>

Tot i aquesta informació, la majoria de les expressions temporals no estan marcades i es troben en el cos del text. Poden fer referència al moment (60a), a la durada (60b) o a la freqüència (60c) d'una entitat clínica o d'un esdeveniment que li succeeix al pacient.

(60)

a. En 2008 se objetivó masa retroperitoneal a raíz de un cuadro de dolor abdominal, con biopsia abdominal que descartó linfoma
b. Refiere presentar desde el 12-13/01 aumento de disnea hasta hacerse de mínimos esfuerzos, así como fiebre de hasta 38.5ºC, ageusia y anòsmia
c. Calcio/vitamina D3 500/400 cada 12 horas

Les expressions temporals poden reflectir diferents tipus de relacions basades en el temps, normalment entre diferents esdeveniments o bé situen un esdeveniment en una línia temporal. A continuació, presentem els tipus de relacions temporals.

c) Les relacions temporals

Les relacions temporals poden ser de diferent tipus segons si s'estableix que una expressió temporal o un esdeveniment són simultanis, posteriors o anteriors respecte d'un altre esdeveniment o d'una altra expressió temporal.

El TimeML, que ja hem presentat a l'apartat 4.2., presenta només un subgrup reduït d'aquestes relacions. Sun et al. (2013a,b), que són la nostra principal referència per a l'anotació de la temporalitat, proposen els següents valors per distingir entre els diferents tipus de relacions temporals:

[5] En aquest apartat (5.3.) els elements subratllats són estructures temporals.

- BEFORE. Aquest valor indica que A és anterior a B.
  (61)

  Contacto previo **a visita**

  En l'exemple (61) 'contacto' és anterior a 'visita': [contacto] BEFORE [a visita].

- AFTER. Aquest valor s'aplica quan A és posterior a B.
  (62)

  Abans de **l'ingrés** refereix **febre**

  En l'exemple (62) 'l'ingrés' és posterior a febre: [l'ingrés] AFTER [febre].

- SIMULTANEOUS. Aquest valor indica que A i B són simultanis.
  (63)

  Fecha de **frotis positivo**: 20/08/2020

  En l'exemple (63) el 'frotis positivo' succeeix el '20/08/2020', de manera que: [frotis positivo] SIMULTANEOUS [20/08/2020].

- OVERLAP. Aquest valor s'aplica quan A i B es superposen. Es tracta generalment de dos esdeveniments coordinats que es relacionen de la mateixa manera amb un altre esdeveniment o una estructura temporal.
  (64)

  A l'ingrés àlgies i **malestar**

  En l'exemple (64) 'àlgies' i 'malestar' són simultanis a 'ingrés', però entre ells es superposen ja que es troben al mateix nivell, de manera que: [àlgies] OVERLAP [malestar].

- BEGUN_BY. Aquest valor fa referència a quan s'especifica l'inici d'un esdeveniment.
  (65)

  Fecha de inicio de **los síntomas**: 06/08/2020

  En aquests cas (65) els 'síntomas' inicien explícitament el '06/08/2020': [síntomas] BEGUN_BY [06/08/2020].

- ENDED_BY. Aquest valor fa referència a quan s'especifica el final d'un esdeveniment.
  (66)

  Se decide retirar **antihistamínicos** a las 17h

  En l'exemple (66) els 'antihistamínicos' es retiren/finalitzen 'a las 17h': [antihistamínicos] ENDED_BY [a las 17h].

- DURING. Aquest valor s'aplica quan A succeeix mentre s'esdevé B.
  (67)

  Se realiza **PCR en urgencias**

  En l'exemple (67), la 'PCR' s'esdevé mentre està 'en urgencias': [PCR] DURING [en urgencias].

Sovint quan s'anoten les relacions temporals es segueix un procés de dues etapes: en una primera fase s'identifiquen les estructures temporals i els esdeveniments que apareixen en el text; en una segona fase, s'estableix la relació temporal, si existeix, entre els esdeveniments que han aparegut en el text o es situen els esdeveniments en el marc d'una expressió temporal.

(68)

Febrícula a las 10 horas del **ingreso**

En l'exemple (68), primerament cal identificar i anotar l'estructura temporal 'a las 10 horas de' i els esdeveniments 'febrícula' i 'ingreso'. A partir d'aquí, podem establir la relació dels dos esdeveniments ('febrícula' i 'ingreso') respecte de l'estructura temporal 'a las 10 horas'. En aquest cas, 'febrícula' és simultani a l'estructura temporal, mentre que 'ingreso' és anterior a l'estructura temporal. Finalment, podem inferir que 'febrícula' és posterior a 'ingreso'.

L'anotació, doncs, de l'oració (68) és:

[febrícula] SIMULTANEOUS [a las 10 horas]
[ingreso] BEFORE [a las 10 horas]
[febrícula] AFTER [ingreso]

5.3.1. Problemes específics de les expressions temporals en el domini mèdic

El nostre tractament de la temporalitat es limita a l'anàlisi d'aquelles estructures temporals que apareixen en expressions de negació i que, per tant, poden interferir i generar ambigüitat a l'hora de detectar el focus de la negació.

A partir de l'anàlisi de les estructures que apareixen al nostre corpus, proposem una classificació en dos grans grups: les estructures temporals de temps absolut i les estructures temporals de temps relatiu.

- Les **estructures de temps absolut** indiquen un moment concret que podem situar fàcilment en una línia temporal. Es tracta de dates de calendari i hores. Són freqüents especialment per indicar el dia i l'hora en què es redacten els informes i les notes clíniques o per especificar el moment en què es du a terme o en què s'ha dut a terme un procediment mèdic:

(69)

a. <date>22.01.2021</date>
   <time>14:22</time>
b. * Antígeno positivo: 18/01.
c. Hemocultivos [30/11]: Negativos
d. <DATE_INI>2020/09/06</DATE_INI>
   <TIME_INI>12:53:50</TIME_INI>
   <DATE_FIN>2020/09/15</DATE_FIN>
   <TIME_FIN>13:10:48</TIME_FIN>

En l'exemple (69a,b,c,d), les estructures temporals que apareixen són de temps absolut ja que es poden situar sense dificultats en una línia temporal. Expressen informació que ens permet saber el moment concret en què succeeix un esdeveniment, és a dir, indiquen el dia, el mes, l'any i, fins i tot, l'hora.

Com existeixen diferents maneres de referir-se a un mateix valor o lapse de temps, la norma ISO 8601 estandarditza les estructures de temps absolut i construeix una fórmula única per a totes, eliminant la variació. Aquesta norma permet detectar que diferents estructures fan referència a un mateix temps, de manera que es pot situar correctament els esdeveniments clínics en una línia temporal.

En els següents exemples extrets del nostre corpus, el *ClUB-21*, observem com s'utilitzen diferents estructures temporals per fer referència a un mateix temps:

(70)

a. <date>*20.08.2020*</date>[6]
<time>22:22</time>
[…]
Fecha de ingreso: *20/08/2020*
Fecha de frotis positivo: *20/08/2020*
Fecha de inicio tratamiento: -

b. 1. Adenocarcinoma de páncreas E-IV (peritoneal, ganglionar). 1L Gem/Abraxane bisemal (C2D8 29/12/20). STOP tratamiento por Covid-19. Feb/21: PE local , ganglionar, peritoneal y hepática (¿en contexto de suspensión de trata miento?)
2. Estreñimiento. Cuadros suboclusivos. Enteritis y colitis colon ascendente/ciego
3. Antecedente de infección reciente por SARS-CoV-2. PCR persiste positiva *(12/02)*
[…]
<date>*15.02.2021*</date>
<time>13:58</time>
[…]
** Ingreso del *12 al 15/02* por sobolcusion intestinal en contesto de CP

En l'exemple (70a), s'expressa de dues maneres diferents la mateixa data: '20.08.2020' i '20/08/2020'. Estandarditzada seguint la norma ISO 8601 s'hauria d'expressar de la següent manera: '2020-08-20'. A més, apareix una hora que estandarditzada s'expressaria així: 'T22:22'.

En l'exemple (70b), s'expressen de diferents maneres dues dates: '12/02' que és igual a '12' i '15.02.2021' que és igual a '15/02'. Aquestes dues dates estandarditzades s'escriurien així:

[6] En l'exemple (70) marquem en cursiva les estructures temporals que fan referència a un mateix temps.

'2021-02-12' i '2021-02-15'. A més també apareix l'estructura temporal '29/12/20' que estandarditzada seria '2020-12-29' i '13:58' que seguint la norma seria 'T13:58'.

- Les **estructures de temps relatiu** no expressen de manera unívoca un moment concret de temps. Com que no expressen explícitament el punt o moment concret en què cal situar l'esdeveniment, la seva interpretació és complexa. Per a la seva interpretació moltes vegades cal recórrer a la informació temporal de les metadades del document, que funcionen com a punts de referència (exemple 69a,d). També poden servir de referència les estructures de temps absolut que apareixen al cos del document.

Les següents estructures són alguns exemples de temps relatiu extrets del nostre corpus:

(71)

a. Avui se li veu milloria del seu estat general
b. No refereix àlgies durant el dia
c. Pendent iniciar Remdesivir avui per la tarda
d. Passa el matí reposant

Com podem veure en les estructures de l'exemple (71), sense un punt de referència explícit no podem situar-les en una línia temporal i, en conseqüència, no sabem el moment exacte en què succeeixen els esdeveniments.

Les estructures de temps relatiu són més freqüents, per la qual cosa el tractament de la temporalitat requereix un procés d'inferència a partir d'aquestes estructures per al correcte ancoratge dels esdeveniments en la línia temporal clínica.

- Una altra problemàtica amb què ens trobem a l'hora de tractar la temporalitat és la detecció de l'esdeveniment que es relaciona temporalment amb l'estructura temporal. Sovint per la puntuació no queda clar on comença i on acaba una oració. A vegades trobem, a més, una mancança de signes de puntuació, cosa que dificulta la interpretació d'allò que s'expressa. Això genera ambigüitat a l'hora d'establir les relacions temporals de les estructures temporals amb els esdeveniments.

(72)

Valorada en enero en el CAP por NRL en 02/2014 derivada tras episodio en enero de disminución del nivel de consciencia por hipotensión arterial (probablemente farmacológica)

En l'exemple (72), podem veure la mancança de signes de puntuació que separin els diferents elements del discurs. Això pot generar ambigüitat a l'hora d'establir les relacions temporals entre les estructures temporals i els esdeveniments. En aquest cas l'estructura 'en 02/2014' pot presentar aquesta problemàtica ja que es troba entre dos esdeveniments i sense signes de puntuació. Això pot generar ambigüitat a l'hora d'establir amb quin esdeveniment té relació i amb quin no.

- També trobem aquesta problemàtica en els enunciats en discurs indirecte. En aquests no queda clar si l'estructura temporal es relaciona amb el *verbum dicendum* o bé amb algun altre esdeveniment.

(73)

No **refereix** àlgies durant el dia

En l'exemple (73) podem observar aquesta ambigüitat. Podem interpretar que allò que succeeix 'durant el dia' és l'esdeveniment 'refereix' o l'esdeveniment 'àlgies'.

6. Proposta d'anotació del focus de la negació en el domini mèdic

A partir de l'estudi de Taulé et al. (2020) i de l'anàlisi de les estructures de negació dels informes mèdics que conformen el corpus *ClUB-21* presentem a continuació una proposta per a l'anotació del focus de la negació en el domini mèdic.

Com Taulé et al. (2020), distingim entre focus explícit i focus implícit i entre construccions sense predicat verbal i construccions amb predicat verbal.

a) Construccions amb el focus de la negació explícit

Els recursos mitjançant els quals es pot explicitar el focus de la negació són: el seu desplaçament, explicitant el subjecte pronominal en cas que sigui el focus, mitjançant construccions contrastives, reforçant la negació generalment amb pronoms o determinants indefinits o mitjançant marcadors tipogràfics.

En el corpus *ClUB-21* trobem poques estructures en què s'utilitzi algun d'aquests recursos per explicitar el focus. El més freqüent són les estructures on es reforça la negació mitjançant pronoms o determinants indefinits, a vegades trobem alguna construcció contrastiva, poques vegades el focus de la negació apareix desplaçat o marcat tipogràficament i no trobem cap exemple on el focus sigui un subjecte pronominal explícit.

D'acord amb Taulé et al. (2020), en les construccions en què trobem un element topicalitzat (74a), complementat per un pronom o un determinant indefinit (74b), en contraposició a una altra construcció (74c) o marcat tipogràficament (74d) considerem que aquest element és el focus de la negació i que està explicitat:

(74)

a. Se realiza [PCR que resulta] **negativa**
b. **No** [se observó *ningún* síntoma]
c. Se entra la cena, *aunque* ***rechaza*** *[la cena]*
d. **No** [presenta *FIEBRE*]

b) Construccions amb el focus de la negació implícit

En les construccions amb el focus implícit no trobem marcadors que ens ajudin en la seva identificació. El reconeixement del focus de la negació es fa mitjançant criteris pragmàtico-semàtics: el context discursiu, l'element més oblic i el significat positiu implícit.

A partir de l'anàlisi de les diferents construccions de negació que apareixen en els informes mèdics, proposem una sistematització per a la detecció del focus de la negació segons el tipus d'estructura negativa. Distingim entre les construccions sense predicat verbal i les construccions amb predicat verbal.

Les **construccions sense predicat verbal** que trobem als informes mèdics són:

i) L'adverbi 'no' seguit d'un substantiu

En aquest tipus de construcció, el substantiu modificat per l'adverbi 'no' és el focus de la negació ja que és l'element directament negat i normalment l'únic element que conforma l'abast.

(75)
a. **No** [algias]
b. **No** [hábitos tóxicos]

ii) L'adverbi 'no' seguit d'un adjectiu o d'un participi

En aquestes estructures l'adjectiu o el participi són el focus de la negació ja que no es nega l'existència del substantiu que precedeix la construcció i al qual es refereixen, sinó que es nega la característica expressada per l'adjectiu.

(76)
a. Pruebas de antígenos **no** [teponémicos]
b. Tos **no** [productiva]

iii) Les construccions en què el marcador de negació és un substantiu o un adjectiu seguits d'una preposició

En aquests casos el focus de la negació és el sintagma nominal introduït per la preposició, el qual sol ser l'únic element de l'abast.

(77)
a. **Remisión de** [la fiebre]
b. Un hombre de 60 años acudió al hospital con una **pérdida repentina de** [sensibilidad]

iv) La preposició 'sense' o 'sin' seguida d'un sintagma nominal o d'un infinitiu

En aquestes estructures el focus de la negació és el sintagma nominal o l'infinitiu que introdueixen, excepte si l'infinitiu té algun complement. Si l'infinitiu té un complement, aquest complement és el focus de la negació (78b).

(78)
a. Mujer de 78 años, **sin** [antecedentes médicos conocidos]
b. La familia de había autoconfinado a causa de la pandemia de COVID-19 durante la semana anterior, **sin** [salir para ir a la escuela ni al trabajo]

v) Construccions amb predicat verbal

Per a les **construccions amb predicat verbal** seguim els criteris de Taulé et al. (2020), que ja hem exposat. En el cas que no hi hagi cap adjunt, si el verb és intransitiu el focus de la negació pot ser el verb o el subjecte (79a), si es tracta d'un verb existencial el focus és el subjecte existencial (79b), si es tracta d'un verb transitiu, copulatiu o que regeix preposició el focus de la negació és el complement directe (79c), l'atribut (79d) o el sintagma preposicional regit (79e)

i si es tracta d'un verb ditransitiu cal aplicar el criteri de l'element més oblic per determinar si el focus de la negació és el complement directe o indirecte (79f).

(79)

a. **Se finaliza** [el tratamiento]
b. **No** [hi ha risc]
c. **No** [palpo masas]
d. **No** [és valorable]
e. **No** [dispone de más información]
f. **No** [le gusta la comida del hospital]

L'exemple (79f) es tracta d'un verb ditransitiu complementat per 'le' i per 'la comida del hospital'. Com el complement directe ('la comida del hospital') aporta més informació, la qual és més prescindible, considerem que és el focus de la negació.

En el cas que aparegui algun adjunt en el predicat verbal, l'adjunt és el focus de la negació ja que es tracta de l'element més oblic.

(80)

a. **No** [presenta edemas en extremidades inferiores]
b. **No** [presenta lesiones en marco óseo]
c. **No** [se inicia el tratamiento por viaje]
d. **No** [toma tratamiento de forma habitual]

Tot i que no és gaire freqüent en els informes mèdics, en el cas que aparegui més d'un adjunt en la mateixa oració, cal aplicar el criteri de l'element més oblic per determinar quin dels adjunts és el focus de la negació.

(81)

Sin [alteraciones en la repolarización en derivaciones precordiales]

vi) Expressions temporals i focus de la negació

A partir de l'anàlisi de les construccions amb predicat verbal que tenen un adjunt que expressa informació temporal, es planteja un problema d'ambigüitat ja que es pot interpretar com a focus de la negació tant l'expressió temporal com un argument o un altre adjunt. Es presenten, doncs, diferents interpretacions possibles: a) el focus és l'estructura temporal aplicant el criteri de l'element més oblic; b) el focus és l'argument o un altre adjunt; c) és una oració amb dos focus, l'argument o adjunt i l'estructura temporal.

Observem la següent oració:

(82)

[Durante este periodo de enfermedad] **no** [tuvo contacto con otros enfermos]

En l'exemple (82) serien possibles dues interpretacions:

INTERPRETACIÓ-1:

Es pot interpretar com 'Tuvo contacto con otros enfermos, pero no durante este periodo de enfermedad', on el focus de la negació seria l'estructura temporal.

INTERPRETACIÓ-2:

Es pot interpretar com 'Durante este periodo de enfermedad tuvo contacto, pero no con otros enfermos', on l'entitat 'con otros enfermos' seria l'element més explícitament negat.

Proposta:

En aquests casos, per norma general, proposem la següent solució: considerem que l'expressió temporal defineix el marc temporal en el qual es realitza el focus de la negació, sigui un argument o un adjunt. És a dir, l'expressió temporal funciona com un delimitador del lapse en què la negació és vàlida. L'argument o adjunt no temporal seria llavors l'element que actuaria com a focus de la negació.

A partir del corpus *ClUB-21*, hem creat un subcorpus format per les oracions en què apareixen estructures temporals en negacions. Aquest subcorpus s'anomena *ClUB-FOC21* i està format per un total de 197 oracions.

vii) Expressions temporals no afectades per la negació

De vegades, en oracions que contenen una negació, les estructures temporals es troben fora de l'abast de l'estructura oracional i funcionen com a marcadors discursius. En aquests casos, l'expressió temporal no es veu afectada per la negació ja que es troba fora de l'abast. Com que el focus de la negació sempre és un element de l'abast, en aquests casos l'estructura temporal no genera ambigüitat a l'hora de detectar el focus ja que no és susceptible de ser interpretat com a tal.

Aquest tipus d'oracions, doncs, no les hem inclòs al *ClUB-FOC21*. És més, considerem que reforcen la nostra hipòtesi: que l'estructura temporal, quan es troba dins de l'abast de la negació, funciona com a delimitador del lapse de temps en què té validesa la negació i no com a focus de la negació. De la mateixa manera, delimita el lapse de temps en què tenen lloc els esdeveniments quan es troba fora de l'abast i és un marcador discursiu.

(83)

a. *Tras valoración* se reinicia ingesta y sedestación, **sin** [incidencias][7]

[7] En les oracions de l'exemple (83) els elements en cursiva són estructures temporals per distinguir-les del focus de la negació (subratllat).

b. Pasa *la tarde* tranquila **sin** [incidencias]
c. *En 2008* se objetivó masa retroperitoneal a raíz de un cuadro de dolor abdominal, con biopsia abdominal que **descartó** [linfoma]

En els exemples (83a,b,c), les estructures temporals 'tras valoración', 'la tarde' i 'en 2008' especifiquen el moment en què tenen lloc els esdeveniments 'se reinicia ingesta y sedestación', 'se objetivó masa retroperitoneal' i 'biopsia abdominal'. Tanmateix també indiquen el moment en què tenen lloc les negacions, és a dir, el moment en què no s'aprecien incidències ('sin incidencias') o en què es descarta el limfoma. Delimiten, per tant, el moment de validesa de la negació. Tanmateix, com no formen part de l'abast de la negació no són susceptibles a ser interpretades com a focus.

viii) Negació marcada morfològicament i expressions temporals

Tampoc hem inclòs en el *ClUB-FOC21* les negacions marcades morfològicament, és a dir, mitjançant un afix. Aquest tipus de marcadors generen una negació interna i considerem, per tant, que no tenen abast atès que no afecten als altres elements oracionals. Com que no tenen abast, si apareix un marcador de negació morfològic junt amb una estructura temporal, aquesta no és susceptible de ser interpretada com a focus de la negació.

(84)
a. Buen estado general, **a**sintomático durante el fin de semana[8]
b. Actualmente **a**sintomático
c. A su llegada a urgencias, **a**febril

Com veiem en els exemples (84a,b,c), l'estructura temporal no pot ser considerada focus de negació, sinó que defineix els límits en què és vàlida la negació generada pel marcador de negació morfològic. Les expressions 'durante el fin de semana', 'actualmente' i 'a su llegada a urgencias' indiquen el lapse en què el pacient no presenta símptomes o febre.

La negació generada mitjançant afixos no afecta, per tant, a les expressions temporals, sinó que nega directament les entitats 'símptomes' i 'febre' i reforça el fet que considerem que, en general, tot i tractar-se d'adjunts, les expressions temporals no són el focus de la negació. El focus serà un argument o bé un altre adjunt.

xix) Casos d'ambigüitat

Trobem expressions temporals que generen ambigüitat a l'hora de detectar el focus de la negació en oracions amb predicat verbal en què l'adjunt temporal complementa el nucli del predicat verbal. El nucli del predicat pot ser un verb negat per l'adverbi 'no' (85a,b,c,d) o bé un verb que funciona com a marcador lèxic de negació (86a,b,c). El verb, a més, sol estar complementat per un argument. A continuació presentem alguns exemples d'aquest tipus de construccions:

[8] En les oracions de l'exemple (84) els elements subratllats són les estructures temporals.

(85)

a. [Ahora] **no** [toma medicación]
b. **No** [realiza deposiciones durante la noche]
c. **No** [refereix àlgies durant el dia]
d. [De momento] **no** [añado enoxaparina]

(86)

a. [19.01.2021] **RETIRADA DE** [CLAVO GAMMA CORTO]
b. [El 20/1] **se suspendió** [myfortic]
c. **Es retira** [Holter aproximadament a les 19:30h]

Aquestes oracions poden suscitar diferents interpretacions:

INTERPRETACIÓ-1:

(85a) "Toma medicación, pero **no** ahora", (85b) "Realiza deposiciones, pero **no** durante la noche", (85c) "Refereix àlgies, però **no** durant el dia", (85d) "Añado enoxoparina, pero **no** de momento".

Seguint aquesta interpretació, el focus de la negació serien les expressions temporals.

INTERPRETACIÓ-2:

(85a) "Ahora toma x, pero **no** medicación", (85b) "Durante la noche realiza x, pero **no** deposiciones", (85c) "Durant el dia refereix x, però **no** àlgies", (85d) "De momento añado x, pero **no** enoxoparina".

D'acord amb aquesta interpretació, el focus de la negació seria un argument oracional.

Proposta:

La nostra proposta consisteix a considerar com a focus de la negació l'argument oracional i l'adjunt oracional com un delimitador del temps en què té lloc la negació. Ens decantem, doncs, per la segona interpretació.

Considerem que els exemples (86a,b,c), on els marcadors de negació són lèxics, reforcen aquesta postura. En aquests casos, com els marcadors són verbs o substantius es veu amb més claredat que neguen directament arguments oracionals.

També es genera aquesta ambigüitat en construccions de negació sense predicat verbal amb un adjunt temporal.

(87)

a. **Sin** [deposiciones desde jueves]
b. **Sin** [pruebas complementarias hoy]
c. [Hemocultivos 30.03]: **negativos**

En aquests casos també es presenten diferents interpretacions. D'acord amb la nostra proposta, les estructures temporals marquen el lapse en què té lloc la negació i el focus de la negació és un altre element oracional. En aquests exemples (87a,b,c) els focus de la negació són: 'deposiciones', 'pruebas complementarias' i 'hemocultivos'.

x) Casos en què l'estructura temporal és el focus de la negació

La nostra proposta de no considerar l'estructura temporal com a focus de la negació és aplicable a aquelles oracions amb arguments o altres adjunts. Trobem, però, algunes oracions en què no hi ha arguments o altres adjunts. En aquests casos considerem que l'estructura temporal sí que és el focus de la negació ja que és l'únic element de l'abast i, per tant, l'únic element susceptible a ser interpretat com a focus.

(88)

a. Reprogramaremos estudio RMN cerebral cuando tenga Acs, **no** [antes de 14 días]
b. **Retirar** [mañana a la misma hora]

7. Línies de futur

El tema del nostre treball va sorgir amb l'objectiu principal de crear una guia d'anotació per al focus de la negació en documents del domini mèdic, atesa la importància d'aquest element en el marc de l'extracció d'informació.

Durant l'anàlisi del problema vam detectar que la temporalitat generava conflictes a l'hora de determinar el focus de la negació. Per aquest motiu el nostre treball es va ampliar i vam incloure la interacció entre temporalitat i negació.

Queden obertes algunes línies de treball que poden abordar-se en el futur:

- Anotar tot el corpus *ClUB-21* ja que per limitacions de temps només s'han anotat una part dels documents.
- Desenvolupar una guia d'anotació complerta per a la negació, on es detallin amb més profunditat aspectes com són l'anotació dels marcadors de negació i de l'abast en el domini mèdic.
- Ampliar l'anotació de la temporalitat: anotar també els esdeveniments, les relacions temporals i les estructures temporals no només en negacions.
- Ampliar el corpus *ClUB-21* per obtenir una mostra més representativa i anotar-lo amb negació (marcadors, abast i focus) i temporalitat.
- Avaluar amb més profunditat l'impacte de les expressions temporals quan ens trobem amb discurs indirecte ja que no està clar a què es refereix l'expressió temporal, si a la negació o al *verbum dicendi*.

8. Conclusions

En aquest treball s'ha realitzat un estudi del subllenguatge mèdic des de la perspectiva de la negació i de la temporalitat. L'objectiu era desenvolupar una guia d'anotació per a l'anotació del focus de la negació en documents del domini mèdic.

S'ha creat el corpus *ClUB-21*, constituït per 24 informes de curs clínic i els seus 24 informes d'alta corresponents.

En el corpus s'han anotat oracions que contenen negació i/o temporalitat d'una part dels documents..

S'ha creat també el corpus *ClUB-FOC21*, constituït per 197 oracions extretes del *ClUB-21* que contenen negació i expressions temporals.

S'ha creat una primera versió de la guia d'anotació del focus de la negació en el domini mèdic. Hem abordat, a més, la problemàtica que presenta l'expressió de la temporalitat en estructures negatives a l'hora de detectar el focus de la negació. S'ha presentat una proposta d'anotació per aquests casos d'ambigüitat.

9. Apèndix

A continuació fem un llistat dels diferents marcadors sintàctics, lèxics i morfològics de negació que trobem al corpus *ClUB-21*.

9.1. Llistat de marcadors de negació del domini mèdic

a) Marcadors sintàctics

- No
- Sense / sin
- No... ni
- Sense / sin... ni
- Mai / nunca

b) Marcadors lèxics

Substantius	*Verbs*	*Adjectius*
- Absència / ausencia	- Acabar	- Absent / ausente
- Desaparició / desaparición	- Deixar / dejar	- Negatiu / negativo
- Finalització / finalización	- Descartar	- Retirat / retirado
- Interrupció / interrupción	- Eliminar	
- Negació / negación	- Finalitzar / finalizar	
- Rebuig / rechazo	- Interrompre / interrumpir	
- Remissió / remisión	- Negar	
- Suspensió / suspensión	- Rebutjar / rechazar	
	- Remetre / remitir	
	- Retirar	
	- Suspendre / suspender	

c) Marcadors morfològics

- Afebril
- Anormal
- Asimptomàtic / asintomático
- Desconegut / desconocido
- Incapacitat / incapacidad
- Inconscient / inconsciente
- Indetectable
- Indolor / indoloro
- Inesperat / inesperado
- Inestabilitat / inestabilidad

- Imprecís / impreciso
- Intolerant / intolerante
- Irregular
- Irrellevant / irrelevante

10. Bibliografia i referències

Alfattni, G., Peek, N., i Nenadic, G. (2020). Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatic*, (108).

Anand, P., i Martell, C. (2012). Annotating the Focus of Negation in terms of Questions Under Discussion. *Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM-2012)*, 65-69.

Banjade, R., i Rus, V. (2016). DT-Neg: Tutorial Dialogues Annotated for Negation Scope and Focus in Context. *European Language Resources Association (ELRA)*, 3768-3771.

Ben Abacha, A., i Zweigenbaum, P. (2011). Medical Entity Recognition: A Comparison of Semantic and Statistical Methods. *Proceedings of the 2011 Workshop on Biomedical Natural Language Processing*, 56-64.

Bethard, S., Palmer, M., Savova, G., i, Pustejovsky, J. (2017). SemEval-2017 Task 12: Clinical TempEval. *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, 565-572.

Blanco, E., i Moldovan, D. (2011). Semantic Representation of Negation Using Focus Detection. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 581-589.

Campillos Llanos, L., Martínez, P., i Segura-Bedmar, I. (2017). A preliminary analysis of negation in Spanish clinical records data set. *Taller de NEGación en ESpañol (NEGES)*.

Chapman, W. W., Bridewell W., Hanbury, P., Cooper, G. F., i Buchanan, B. G. (2001). A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5), 301-310.

Cotik, V., Filippo, D., Roller, R., Uszkoreit, H., i Xu, F. (2017). Creation of an Annotated Corpus of Spanish Radiology Reports. *Proceedings of WiNLP 2017*.

Cruz Díaz, N. P., Morante, R., Maña, M. J., Mata, J., i Parra, C. L. (2017). Annotating Negation in Spanish Clinical Texts. *Proceedings of the Workshop Computational Semantics Beyond Events and Roles (SemBEaR)*, 53-58.

Delàs, J. (2005). Informes clínics, eines de comunicació. *Quaderns de la Bona Praxi*, (18).

Estopà, R. (2020). L'informe mèdic: com millorar-ne la redacció per facilitar-ne la comprensió. *Quaderns 47*.

Galescu, L., i Blaylock, N. (2012). A Corpus of Clinical Narratives Annotated with Temporal Information. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*.

Huddleston, R., i Pullum, G. (2002). *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

Jiménez-Zafra, S. M., Morante, R., Martín-Valdivia, M. T., i Ureña-López, L. A. (2020). Corpora Annotated with Negation: An Overview. *Computational Linguistics*, 46(1), 190-244.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., i Taboada, M. (2019). The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 1–36.

Lima, S., Perez, N., Cuadros, M., i Rigau, G. (2020). NUBES: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 5772-5781.

Llorens, H., Navarro, B., i Saquete, E. (2009). Detección de Expresiones Temporales TimeML en Catalán mediante Roles Semánticos y Redes Semánticas. *Procesamiento del Lenguaje Natural*, (43), 13-21.

Marimon, M., Vivaldi, J., i Bel, N. (2017). Annotation of negation in the IULA Spanish Clinical Record Corpus. *Proceedings of the Workshop Computational Semantics Beyond Events and Roles (SemBEaR)*, 43-52.

Morante, R., i Blanco, E. (2012). *SEM 2012 Shared Task: Resolving the Scope and Focus of Negation. *Association for Computational Linguistics*, 265-274.

Morante, R., i Sporleder, C. (2012). Modality and Negation: An Introduction to the Special Issue. *Computational Linguistics*, (38), 223-260.

Mutalik, P. G., Deshpande, A., i Nadkarni, P. M. (2001). Use of general-purpose negation detection to augment concept indexing of Medical documents: A quantitative study using the UMLS. *Journal of the American Medical Informatics Association*, 8(6), 598–609.

Nadeau, D., i Sekine, S., (2007). A survey of named entity recognition and classification. *In journal of linguistic investigations*, 30(1), 3-26.

Oronoz, M., Gojenola, K., Pérez, A., Díaz de Ilarraza, A., i Casillas, A. (2015). On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56, 318-332.

Palmer, M., Gildea, D., i Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 71–106.

Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., i Katz, G. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. *IWCS-5 Fifth International Workshop on Computational Semantics*.

Pustejovsky, J., Knippen, B., Littman, J., i Saurí, R. (2005). Temporal and event information in natural language text. *Language Resources and Evaluation*, 39(2), 123–164.

Pustejovsky, J., Lee, K., Bunt, H., i Rosmary, L. (2010). ISO-TimeML: An International Standard for Semantic Annotation. *LREC*, (10), 394–397.

Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., i Pustejovsky, J. (2006). TimeML Annotation Guidelines Version 1.2.1. https://www.researchgate.net/publication/248737128_TimeML_Annotation_Guidelines_Version_121

Saurí, R., Saquete, E., i Pustejovsky, J. (2010). *Annotating Time Expressions in Spanish. TimeML Annotation Guidelines. Version TempEval-2010*. Barcelona Media Technical Report BM 2010-02.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., i Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102-107.

Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., C de Groen, P., Erickson, B., Miller, T., Savova, G., i Pustejovsky, J. (2014). Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 143-154.

Sun, W., Rumshisky, A., i Uzuner, O. (2013a). Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, S5-S12.

Sun, W., Rumshisky, A., i Uzuner, O. (2013b). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc*, (20), 806-813.

Taulé, M., Nofre, M., González, M., i Martí, M. A. (2020). Focus of negation: Its identification in Spanish. *Natural Language Engineering*, 1-22.

Verhagen, M., Mani, I., Saurí, R., Knippen, R., Bae Jang, S., Littman, J., Rumshisky, A., Phillips, J., i Pustejovsky, J. (2005). Automating Temporal Annotation with TARSQI. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 81-84.

Wible, D., i Nai-Lung, T. (2010). StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions. *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, 25-31.