UNIVERSITAT DE BARCELONA

Grau de Lingüística

Treball Fi de Grau

Curs 2024-2025

Avaluació i millora de la terminologia del sistema SCRIBAL en l'entorn universitari

Clara Puigventós Martínez

Tutora: Mireia Farrús Cabeceran

Barcelona, 19 de juny de 2025

UNIVERSITAT DE BARCELONA

Coordinació d'Estudis
Facultat de Filologia

Gran Via
de les Corts Catalanes, 585
08007 Barcelona

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

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

Barcelona, a 19 de juny de 2025

Signatura:

PUIGVENTOS MARTINEZ CLARA - 39987912C

Firmado digitalmente por PUIGVENTOS MARTINEZ CLARA - 39987912C
Fecha: 2025.06.19 14:42:03 +02'00'

Resum

SCRIBAL és un sistema de transcripció i traducció automàtica en temps real que facilita l'accessibilitat a l'educació superior, especialment per a estudiants amb discapacitats auditives o que no dominen la llengua de l'ensenyament. L'objectiu principal d'aquest treball ha estat optimitzar-ne el rendiment en l'àmbit del vocabulari específic, concretament el de la medicina. A partir d'una avaluació inicial amb les mètriques *Word Error Rate* i *Character Error Rate*, s'ha detectat que, tot i una bona actuació general, el sistema presenta mancances en terminologia especialitzada. Per resoldre-ho, s'ha entrenat el model amb un corpus sintètic generat amb tecnologies *Text-To-Speech*. El model entrenat ha reduït significativament els errors en contextos específics, tot i que hi ha hagut un lleu empitjorament en contextos generals. Finalment, s'ha dut a terme una agrupació de graus per afinitat lèxica amb l'objectiu d'optimitzar futures adaptacions. Aquest estudi evidencia el potencial d'SCRIBAL com a eina inclusiva en l'àmbit educatiu.

Paraules clau: lingüística computacional, reconeixement automàtic de la parla, tecnologies TTS, adaptació terminològica, SCRIBAL.

Abstract

SCRIBAL is a real-time automatic transcription and translation system that facilitates accessibility to higher education, particularly for students with hearing impairments or who don't have full command of the language of instruction. The main goal of this study has been to optimize its performance in the area of specialized vocabulary, specifically in medicine. Based on an initial evaluation using the Word Error Rate and Character Error Rate metrics, it was found that, despite generally good performance, the system shows limitations in handling specialized terminology. To address this issue, the model was trained with a synthetic corpus generated using Text-To-Speech technologies. The fine-tuned model significantly reduced errors in specific contexts, although there was a slight decline in performance in general contexts. Finally, a grouping of degree programs by lexical similarity was carried out to support more efficient future adaptations. This study demonstrates SCRIBAL's potential as an inclusive tool in educational settings.

Keywords: computational linguistics, automatic speech recognition, TTS technologies, terminology adaptation, SCRIBAL.

«Va pensar que l'animaria molt tenir un company, encara que fos una màquina.»

Faules de robots, Stanisław Lem

Aquest treball no hauria estat possible sense la confiança que m'ha donat la meva tutora, Mireia Farrús, ni l'ajuda dels tècnics de l'SCRIBAL Pol i Mauro. A ells, i a la resta de membres de l'equip, moltes gràcies.

ÍNDEX

ÍNDEX DE FIGURES

ÍNDEX DE TAULES

1 INTRODUCCIÓ

Les tecnologies de la parla tenen l'objectiu de fer possible la interacció entre les persones i els ordinadors a través de la llengua oral (Llisterri, 2009). Permeten fer de manera més senzilla i ràpida algunes tasques que involucren l'escriptura o la lectura, com per exemple apuntar recordatoris o interactuar amb contestadors automàtics. Una de les aplicacions de les tecnologies de la parla més coneguda és la síntesi de la parla, és a dir, produir veu de forma artificial, per la qual al segle XVIII ja s'havien creat els primers models —per exemple el de von Kempelen (Jurafsky and Martin, 2025a, p. 26). Tanmateix, una segona aplicació de les tecnologies va començar a agafar embranzida al segle XX: el reconeixement automàtic de la parla (ASR). Des de "Radio Rex", una joguina dels anys 20 que es movia en sentir el seu nom, fins als sistemes que es troben integrats avui en dia als telèfons mòbils o en els assistents personals, l'ASR ha esdevingut una eina més en el dia a dia de moltes persones. A més, aquest sistema, que principalment es tradueix com a transcripció automàtica, ha resultat útil en entorns en què els usuaris tenen algun tipus de discapacitat: en el cas de persones sordes, tenir subtítols automatitzats a diferents plataformes els ha permès gaudir d'accessibilitat a continguts pensats, generalment, per a oients.

Més enllà dels beneficis individuals que poden comportar aquests sistemes, és important pensar què poden aportar a altres àmbits més enllà de l'entreteniment. En medicina, per exemple, l'ús de transcriptors automàtics per escriure informes mèdics pot alliberar de feina els metges, els quals es poden dedicar més als seus pacients. Un dels camps en què els ASR poden resultar potencialment útils és el de la docència. Un transcriptor automàtic en temps real de les classes facilitaria l'accessibilitat als seus continguts a estudiants amb discapacitat auditiva o nouvinguts amb una llengua diferent de la de l'ensenyament (Murphy and Hallinger (1985), Kulkarni et al. (2024)). Si a més s'hi incorporés un traductor automàtic, una altra aplicació dins d'aquest domini es trobaria a les aules d'acollida, on els estudiants podrien accedir al contingut de la classe en la seva llengua d'origen i, així, facilitar-los l'aprenentatge i la integració.

L'ús dels ASR per a l'educació i l'ensenyament s'ha explorat des de diversos enfocaments. Des de crear transcriptors per a classes i seminaris (Lamel et al. (2005), Bell et al. (2013)), fins a la incorporació de traductors automàtics a conferències (Fügen

et al., 2007) i classes amb estudiants estrangers que no dominen la llengua d'impartició (Müller et al., 2016). Si bé aquestes propostes s'han fet per a promoure l'ús de llengües diferents de l'anglès, com l'alemany en el cas de Müller et al., de moment no hi ha cap transcriptor i traductor automàtic en temps real en català pensat específicament per a la docència. És per això que, des del grup de recerca Centre de Llenguatge i Computació (CLiC) de la Universitat de Barcelona, s'està desenvolupant l'SCRIBAL, un escriba digital que pugui fer front a aquestes necessitats.

El projecte SCRIBAL[1] és un servei de transcripció i traducció automàtica en temps real per a les institucions educatives que té l'objectiu de transcriure de forma ràpida i precisa classes, conferències i altres continguts acadèmics en àudio. Aquest servei està pensat per facilitar l'accessibilitat a persones amb discapacitats auditives als continguts acadèmics, però també a aquelles persones que no tenen un bon domini de la llengua en què s'imparteixen les classes, normalment el català.

A diferència d'altres sistemes, SCRIBAL permet la transcripció —i si cal la traducció— del contingut oral de les classes o conferències, mitjançant la subtitulació de presentacions o directament a l'ordinador o mòbil de l'alumne. A més, el sistema pot ser autogestionat per la universitat, cosa que proporciona més control sobre la privacitat i la seguretat de les dades. Els trets distintius d'aquest sistema són l'adaptació als dialectes del català i a la terminologia específica de les universitats.

El que es vol aconseguir amb aquest projecte és crear un impacte positiu en la societat i facilitar l'accessibilitat a l'educació superior per a tots els estudiants, independentment de les seves capacitats auditives o del seu nivell de competència en la llengua d'instrucció.

Amb aquest treball s'ha contribuït a fer l'adaptació terminològica de l'SCRIBAL per al camp de la medicina, a través d'un refinament amb dades sintètiques i d'un test amb dades reals, amb l'esperança que, en un futur, es puguin continuar millorant la resta d'especialitats universitàries. També s'ha fet una agrupació de les branques del coneixement segons el seu vocabulari específic per veure si un mateix vocabulari pot

[1] Ajut producte del programa indústria del coneixement finançat per l'AGAUR (ref. PROD 00016). Pàgina web: scribal.cat

servir d'entrenament per a altres camps lingüístics.

Aquest treball es divideix en tres parts: una primera part d'avaluació del model —general per a totes les branques del coneixement i específica per a la branca seleccionada per millorar—, una segona part d'entrenament i refinament del model —amb la seva avaluació— i una proposta d'agrupació per branques de coneixement. Cada una s'explica en apartats separats.

2 FONAMENTS TEÒRICS

2.1 El reconeixement automàtic de la parla

El reconeixement automàtic de la parla, *automatic speech recognition* en anglès, és una branca de les tecnologies de la parla que ha anat adquirint, en els últims anys, rellevància en les tasques quotidianes. La transcripció automàtica és l'aplicació més important de l'ASR i es pot trobar en la generació automàtica de subtítols en àudios i vídeos d'algunes aplicacions o en feines on el dictat és primordial, per exemple en l'àmbit del dret (Jurafsky and Martin, 2025a, p. 1).

Tanmateix, malgrat els avenços per aconseguir unes transcripcions cada cop més precises i similars a les que faria un ésser humà, encara queda molt camí per recórrer. Cada àmbit en què s'utilitzen té necessitats diferents, fet que demana una adaptació específica dels models per a les tasques que han de resoldre. Jurafsky i Martin (2025a, p. 2-3) exposen quatre tipus de variació que s'ha de valorar a l'hora de crear o adaptar un model de transcripció automàtica:

- En primer lloc, hi ha la variació de la **mida del vocabulari**. Hi ha tasques de reconeixement de la parla en què es necessita un vocabulari molt reduït, com els contestadors automàtics que reconeixen dígits o les paraules *sí* i *no*. Aquest tipus de tasques acostumen a tenir pocs errors a causa de la poca variació de mots que cal reconèixer, mentre que la transcripció de converses o vídeos requereix l'ús de bases de dades molt més àmplies. La presència de paraules similars o de múltiples opcions en un mateix context fa que sigui una tasca molt més complicada de dur a terme.

- En segon lloc, hi ha la variació de l'**interlocutor**. Els discursos preparats, la lectura en veu alta o un humà parlant a una màquina són converses més fàcils d'identificar a causa de la claredat de la dicció, que és més acurada, i el ritme en què es parla, que és més baix. En canvi, el reconeixement de la conversa entre humans resulta ser una tasca molt més complicada per als ASR, en què els interlocutors parlen més de pressa, no acaben les frases i no vocalitzen amb claredat.

- En tercer lloc, es parla de la variació del **canal i el soroll**. Un micròfon a prop

del parlant en un ambient tranquil farà que l'àudio es processi més fàcilment que no pas si l'enregistrament té lloc en un espai sorollós i amb interferències constants que en dificultin el reconeixement.

- Finalment, hi ha la variació del **parlant**. Els models de reconeixement de la parla s'acostumen a entrenar amb un set de dades limitat; normalment d'una mateixa regió i de la mateixa franja d'edat. Això fa que algunes varietats dialectals o la parla dels nens petits es vegi infrarepresentada i, per tant, no sigui reconeguda pels ASR.

2.1.1 Necessitats de l'SCRIBAL

L'SCRIBAL, en ser un transcriptor per a l'àmbit acadèmic universitari, s'ha d'adaptar a les variacions més complexes dels reconeixedors automàtics de la parla:

- Ha de tenir una gran base de dades per adaptar-se a tots els àmbits acadèmics.
- Ha de ser capaç de reconèixer els discursos d'un professor amb les possibles intervencions espontànies dels alumnes i els debats que es puguin generar a l'aula.
- Ha d'entrenar-se en un ambient *semisorollós*, amb la possibilitat que els professors i els alumnes no disposin de micròfon propi i que hi hagi interferències.
- Ha de reconèixer els diferents tipus de parla, accents i dialectes d'una gran quantitat de professors, així com dels alumnes que facin intervencions.

Tot i que el model té en compte aquestes variacions, se centra especialment en la primera i l'última. La primera perquè busca ampliar el vocabulari per tal que hi hagi una millor actuació en àrees on la terminologia específica no s'ha entrenat prou. La segona perquè es vol crear un transcriptor i traductor capaç de fer-se servir en qualsevol entorn dels Països Catalans, de manera que cal que entengui tots els dialectes del català.

2.1.2 Large Language Models

Generalment, els models d'ASR funcionen mitjançant una arquitectura *encoder-decoder*. Aquesta arquitectura consisteix en la vectorització d'un segment de la parla —normalment un so— al qual se li assigna la seva lletra corresponent, de manera que unes

característiques acústiques s'associen a una grafia (Jurafsky and Martin, 2025a, p. 10). D'aquesta manera, quan un so arriba a un transcriptor, aquest li associa el text corresponent. De la mateixa manera funcionen els sintetitzadors de la parla, amb la diferència que és a partir del text que reben que produeixen artificialment els segments acústics que tenen assignats. Aquesta aproximació, però, s'ha vist millorada gràcies als models de llenguatge extensos —*Large Language Models* (LLMs) en anglès.

Els LLMs són models de llenguatge que s'entrenen amb grans quantitats de text de les quals aprenen sobre les estructures lingüístiques i el món. Gràcies a aquest aprenentatge són capaços de resoldre les tasques de llenguatge natural —entre les quals les anteriorment esmentades— de forma excepcional (Jurafsky and Martin, 2025b, p. 1). Això és gràcies a la quantitat d'exemples de llenguatge natural que conformen els seus corpus d'entrenament.

Un dels models més destacats els últims anys és Whisper d'Open AI (OpenAI, 2022), basat en xarxes neuronals. Aquest sistema d'ASR és de codi lliure, fet que permet que es pugui millorar i adaptar fàcilment a les necessitats de qui l'empri. Per exemple, es poden crear filtres que evitin certes paraules ofensives, innecessàries o mal transcrites. Des de plataformes com Softcatalà s'ha fet servir el model Whisper —amb el motor `CTranslate2`— per millorar la seva actuació en català (Softcatalà, 2023). SCRIBAL adopta aquest model base i l'adapta al seu context.

Whisper és pioner pel que fa al reconeixement de la parla. Ha estat entrenat amb un corpus de 680.000 hores de parla multilingüe anotades i supervisades, cosa que el fa un model robust i que millora en la fiabilitat davant dels accents, el soroll de fons i el llenguatge tècnic (Radford et al., 2022). En ser un corpus multilingüe, permet, a més de la transcripció, una traducció fiable, sobretot a l'anglès.

Amb tot, si bé aquest model s'ha entrenat amb moltes hores, la majoria de les dades són en anglès; només el 17 % (117.113 hores) correspon a dades en altres llengües. Pel que fa a les dades corresponents a traducció de textos en anglès, representen el 18 % (125,739 hores) del total. Concretament, el català ha format part d'aquesta mostra de llengües: amb 1883 hores per al reconeixement automàtic de la parla i amb 236 hores per a la traducció (Radford et al., 2022, p. 27). Aquestes hores del català, tot i ser una quantitat important respecte d'altres llengües, no són suficients per garantir una

transcripció precisa en tots els àmbits del coneixement.

De fet, malgrat la bona actuació d'aquests sistemes en situacions quotidianes, no se'n surten tan bé en contextos de camp específic (Pawlowicz et al., 2025). Per exemple, en medicina és difícil trobar bases de dades amb la terminologia específica d'aquesta branca per poder-los entrenar (Huh et al. (2023), Maulik et al. (2024)), de manera que cal buscar o crear noves maneres de refinar els models perquè tinguin una bona actuació en aquelles àrees de què disposen de menys dades.

2.2 L'adaptació terminològica

L'aproximació més comuna a l'hora de refinar —*fine-tune*— els LLMs per a terminologia específica consisteix a buscar o crear una base de dades anotada amb el vocabulari que s'hi vol incorporar. Una part d'aquesta base de dades es fa servir per entrenar el model i una altra part per avaluar-lo. Aquesta metodologia és la que es troba a Pawlowicz et al. (2025) i Maulik et al. (2024). Tanmateix, hi ha altres metodologies que s'estan explorant.

Una d'elles és la multimodalitat: Huh et al. (2023) proposen entrenar un model general d'ASR fent servir el mètode *Vision Language Pre-training*, que consisteix a combinar informació textual i visual per adaptar les transcripcions a partir del coneixement que aporta la imatge.

Una altra aproximació és la de Lall and Liu (2024) que proposen millorar el vocabulari de camp específic a partir del biaix contextual. Això és, crear una petita base de dades amb el vocabulari necessari i dirigir-hi l'*output* del model, per tal de guiar la transcripció final.

En darrer terme trobem els models de *fine-tuning* anotats a través de sintetitzadors de la parla —sistemes *Text-to-Speech* (TTS). Aconseguir dades anotades i supervisades a vegades resulta difícil i costós, i en molts casos a vegades només es té accés a corpus textuals. Diferents propostes (Vásquez-Correa et al. (2023), Joshi and Singh (2022), Tran et al. (2025)) exploren la possibilitat de fer servir sistemes TTS per crear dades sintètiques i anotades a partir de corpus textuals, fet que resulta amb una reducció de costos i una millora dels models. En català es poden trobar models TTS, com *Matxa-*

TTS[2], del Barcelona Supercomputing Center (BSC). Aquest sistema de codi obert està disponible en quatre variants dialectals —central, nord-occidental, balear i valencià— i amb vuit parlants diferents, tant homes com dones (Barcelona Supercomputing Center, 2024).

2.3 Mètriques d'avaluació

Tot i la diversitat d'aproximacions pel que fa a l'entrenament, l'avaluació dels models de transcripció és, en general, universal. El *word error rate* (WER) és la mètrica d'avaluació estàndard per a sistemes de reconeixement de la parla. Es basa en la diferència entre una cadena de paraules de referència —*Gold-Standard* (GS)— i la mateixa cadena de paraules produïda pel transcriptor. Per trobar aquesta diferència, cal calcular la distància mínima entre paraules d'ambdues cadenes. Això ofereix el nombre mínim de paraules substituïdes, inserides i eliminades respecte del Gold-Standard (Jurafsky and Martin, 2025a, p. 16). La WER es defineix de la següent manera[3]:

$$\text{Word Error Rate [\%]} = 100 \times \frac{\text{Insercions} + \text{Substitucions} + \text{Elisions}}{\text{Paraules totals del } \textit{Gold-Standard}} \tag{1}$$

A continuació es mostra un exemple d'una frase (GS) amb la seva possible transcripció (TR) i els tipus d'errors que troba la WER (Aval) —dues insercions, una elisió i una substitució—, així com el càlcul del percentatge d'error:

GS:	M'agraden	***	***	les	panses	i	els	pinyons.
TR:	M'agraden	les	prunes	les	panses		els	pinyols
Aval:		I	I			E		S

$$\text{Word Error Rate} = 100 \times \frac{2 + 1 + 1}{8} = 50\% \tag{2}$$

Si la WER resulta amb un percentatge elevat, la transcripció que ha fet el model és poc acurada. Com més baixa és la WER, més aproximada serà la transcripció al *Gold-*

[2]Basat en l'arquitectura Matcha-TTS i Vocos (Mehta et al., 2024).

[3]Com que l'equació inclou insercions, els resultats poden ser superiors a 100 %.

Standard. És difícil no aconseguir cap error, ja que fins i tot possibles transcripcions que un humà validaria com a correctes poden diferir del text proposat. És per això que un percentatge d'entre 5 i 10 % ja es considera una molt bona transcripció Eric-Urban (2025). A partir del 20 %, encara es considera acceptable, però amb perspectives de seguir entrenant el model. Un error superior al 30 % significa que el model s'ha d'entrenar millor i la transcripció es considera inacceptable.

De manera anàloga a la WER es poden establir altres mètriques segons les necessitats avaluatives de cada model. El *Character Error Rate* (CER) fa el còmput dels caràcters correctes en una cadena de text; el *Sentence Error Rate* (SER) calcula el percentatge de les frases que tenen almenys una paraula errònia.

3 METODOLOGIA

Per tal d'implementar les millores terminològiques específiques al model base de l'SCRIBAL, calia veure, primer, quines eren les seves mancances, tan generals com en els àmbits específics de cada branca universitària. L'avaluació inicial, doncs, tenia dos objectius:

1. Identificar els errors més freqüents del model i classificar-los.
2. Detectar en quina àrea del coneixement es produïen més errors terminològics.

Aquests punts de partida permetrien, per una banda, detectar de forma precoç aquells errors no terminològics i corregir-los a partir de canvis directes al model —per exemple, fent servir filtres—; i, per altra banda, detectar en quins àmbits terminològics específics el model fallava més per dedicar-li les millores immediates.

3.1 Anàlisi inicial del model

Per començar a fer l'avaluació calia buscar bases de dades que continguessin vocabulari universitari en llengua catalana. Els Serveis Lingüístics de la Universitat de Barcelona ens van proporcionar llistes de vocabulari específic d'alguns graus de la UB i els plans docents de diferents assignatures de tretze facultats de la universitat. Els documents que van facilitar constaven de:

- 52 descripcions de graus i màsters amb la llista de totes les assignatures que s'hi cursen (assignatures)[4].
- 1.185 plans docents de diferents assignatures de totes les àrees del coneixement (PD).
- 53 llistes de vocabulari en català, castellà i anglès, algunes de les quals també contenien les definicions dels elements, de diferents branques (UBterm).

[4]El nom que hi ha entre parèntesis és el que es farà servir a partir d'ara per fer referència a cada una de les bases de dades proporcionades pels Serveis Lingüístics.

Aquestes dades es van fer servir per avaluar i entrenar el model durant les diferents fases del projecte. D'una banda, les descripcions de les assignatures i les dades dels plans docents van permetre fer una primera detecció dels errors de transcripció i unes anàlisis qualitativa i quantitativa. De l'altra, amb les llistes de vocabulari es va fer una agrupació per branques de coneixement, per així dibuixar una idea general sobre quins graus podien estar relacionats terminològicament.

Un cop feta l'avaluació general i detectada l'àrea de coneixement que requerís més entrenament, es va fer una avaluació específica d'aquesta àrea.

3.1.1 Avaluació general

Per fer la primera avaluació es va crear un corpus amb la base de dades de les *assignatures*, que es va llegir i enregistrar amb el micròfon d'un ordinador portàtil. Per simular un entorn de parla similar al que hi hauria en una classe, no es va fer servir un micròfon d'alta qualitat. Per a cada grau es va crear un fitxer de text, amb les assignatures llegides i va rebre el mateix nom que el del seu fitxer d'àudio. D'aquesta manera, es va obtenir un primer corpus de treball anotat d'aproximadament 1:34h —anomenat *Corpus Assignatures*— amb dades de totes les facultats que s'havien proporcionat.

Després, es van passar els àudios enregistrats pel model base d'SCRIBAL (`whisper-ctranslate2 -language ca -model large-v3-turbo`) i se'n va extreure la transcripció automàtica. Amb aquesta transcripció i el text original es van fer dues anàlisis comparatives: la quantitativa i la qualitativa.

Per a l'anàlisi **quantitativa**, es van implementar les mètriques WER i CER —normalitzades[5] perquè els espais i les majúscules no influïssin en els resultats— per veure quins textos transcrits havien tingut més errors respecte de l'original i quins menys. Aquestes dades es van traspassar en dos fulls de càlcul per poder ser analitzades, i des dels quals es van crear taules dinàmiques per disposar els resultats visualment segons l'error dels graus de cada facultat.

Per a l'anàlisi **qualitativa**, es van comparar manualment tots els fitxers de text, els

[5]La normalització d'un text és el procés d'adaptació d'un text per a ser comparat amb un altre. Aquest procés pot implicar l'eliminació de majúscules, dígits, simplificació dels signes de puntuació o abreviacions i, en general, tots aquells elements que poden dificultar la seva comparació.

originals amb la seva transcripció, per veure quins tipus d'errors s'havien comès. A partir dels resultats, es van anar creant diferents categories d'error per classificar-los. Els resultats es van anotar en un document d'Excel, on es va calcular el total d'errors per facultat i tipologia. Això també va permetre crear diferents taules per visualitzar les dades. Finalment, es van anotar els diferents tipus d'errors terminològics.

3.1.2 Avaluació específica

Un cop detectats els errors principals del model i la branca de coneixement que obtenia menys precisió, se'n va fer una avaluació específica. La familiarització prèvia amb les bases de dades va permetre adaptar la normalització del text al tipus de corpus que es tenia, a part de resoldre els errors tècnics del model. Això volia dir:

- Canviar els números romans per números àrabs.
- Treure tots els signes de puntuació no reconeixibles pel model —punts i coma, guions i parèntesis— i substituir-los per una coma, ja que la pausa entonativa de tots aquests elements és similar.
- Aplicar un filtre que impedís les confabulacions en els silencis dels àudios (`whisper-ctranslate2 -language ca -model large-v3-turbo -vad_filter`).
- Segmentar els àudios en 30s per facilitar la transcripció completa al model i evitar les omissions.

Aquestes correccions es van afegir a la implementació de les WER i CER utilitzades prèviament.

Per implementar el nou codi a les dades de medicina, calia crear un nou corpus — *Corpus Medicina* (53 min 27 s). Es van agafar els PD i l'UBterm[6] de medicina i es van enregistrar en àudio. El pla docent del grau en odontologia era molt més extens que els altres, així que, per no esbiaixar els resultats, es va escurçar de manera que hi hagués una quantitat similar d'àudios per a tots els plans docents. En aquest corpus s'hi van afegir les dades del Corpus Assignatures corresponents a la Facultat de Medicina. Els

[6]El vocabulari UBterm de medicina constava de l'assignatura *Anatomia Patològica*.

enregistraments es van transcriure amb el model SCRIBAL amb els filtres i després es van comparar amb els seus textos de referència fent servir les mètriques WER i CER adaptades. Posteriorment es va fer una anàlisi qualitativa.

3.2 Entrenament del model

L'entrenament del model d'SCRIBAL es va fer a través d'un corpus creat amb veus sintètiques. A partir dels textos del Corpus Medicina, es van utilitzar les vuit veus del model Matxa-TTS del BSC amb el paràmetre `lenght_scale`, que regula la velocitat a 0,7, 0,85 i 1, per a crear un total de 24 veus sintètiques per a cada text[7].

Un cop preparades les dades, es va seguir la metodologia de *fine-tuning* de Hu et al. (2022), fent servir una GPU RTX 4090 amb 24G de VRAM. El model es va entrenar per 4 èpoques amb un *batch size* de 64. Els paràmetres de LoRa eren $r = 32$ i $\alpha = 64$.

Finalment, es va avaluar el model amb els enregistraments del Corpus Medicina, per veure la millora del vocabulari específic, i amb corpus generals, per veure si afectava la resta de l'actuació del model en altres àmbits.

3.3 Agrupació per branques de coneixement

Amb la base de dades UBterm, a la qual hi havia llistes de vocabulari de 53 graus diferents, es va dibuixar un esquema de les branques de coneixement amb més vocabulari comú. Això serviria per veure si es podia agrupar el vocabulari de diferents àrees per crear corpus més grans i especialitzats per a cada una d'elles.

Per fer això es va crear un fitxer de text per a cada grau amb únicament les paraules en català. Tampoc es va afegir la definició d'aquells graus que la tenien. Amb l'ajuda del codi de Python *Dendrograma*, es van comparar les paraules de cada fitxer de text, per veure quins graus tenien més paraules en comú i, per tant, veure quins es podrien agrupar per àmbits de coneixement. A partir de les comparacions individuals, es van crear clústers amb els graus que compartien més vocabulari i en van sortir 22 grups (Apèndix C). Posteriorment, es va crear un dendrograma que permetés relacionar aquests grups segons la distància que hi havia entre les paraules que tenien en comú.

[7]El codi per generar les veus es pot trobar aquí.

4 RESULTATS

4.1 Avaluació general

4.1.1 Anàlisi quantitativa

L'anàlisi quantitativa es va fer gràcies a les mètriques WER i CER.

Tenint en compte aquests aspectes, els resultats de la CER del model base d'SCRIBAL (presents a la taula 1, on es veuen els valors més importants resumits [8]) ens indiquen que la gran majoria dels graus s'han transcrit correctament o amb errors mínims —51 dels 49 totals. Fins i tot n'hi ha un parell, —Direcció i gestió de centres educatius i Geografia i canvi global— que s'han transcrit sense cap error. Una explicació d'aquest resultat podria ser que la llargada del text i l'àudio d'aquests dos màsters era relativament més curta en relació amb els graus i màsters que han obtingut una taxa CER més elevada.

% de qualitat	CER	Grau o màster	Total
> 30 %	48,61	seguretat-alimentària	3
	45,53	química	
	33,70	ecologia-gestió-restauració	
20-30 %	27,75	gestió-informació-documentació-digital	3
	24,98	ciències-ambientals	
	23,55	estadística	
< 20 %	0,34	lògica-pura-aplicada	51
	0	direcció-gestió-centres-educatius	
	0	geografia-canvi-global	

Taula 1: Resum dels valors CER del model base més rellevants.

Tanmateix, hi ha tres graus i màsters —Gestió de la informació i la documentació digital, Ciències ambientals i Estadística— que es troben entre el 20-30 % d'error, fet que indica que cal implementar-hi millores i més entrenament per acabar de refinar-los. A més, hi ha tres graus i màsters —Seguretat alimentària, Química i Ecologia, gestió

[8] Es pot veure la taula completa a l'Apèndix A, taula 7.

i restauració— amb una CER més alta de 30, que indica una transcripció pobra i una necessitat d'actuació i entrenament més profunds.

Els resultats són molt similars si es valora la WER. Tot i que la majoria de graus encara es troben en el rang de bona transcripció, es veu que els valors (taula 2) augmenten en les tres categories[9].

% de qualitat	WER	Grau o màster	Total
	67,72	química	
> 30 %	47,43	seguretat-alimentària	15
	46,86	conservació-restauració-béns-culturals-tg1108	
	29,41	sociologia	
20-30 %	29,16	direcció-estratègica-seguretat	13
	27,51	ciències-ambientals	
	1,61	psicogerontologia	
< 20 %	0	direcció-gestió-centres-educatius	29
	0	geografia-canvi-global	

Taula 2: Resum dels valors WER del model base més rellevants.

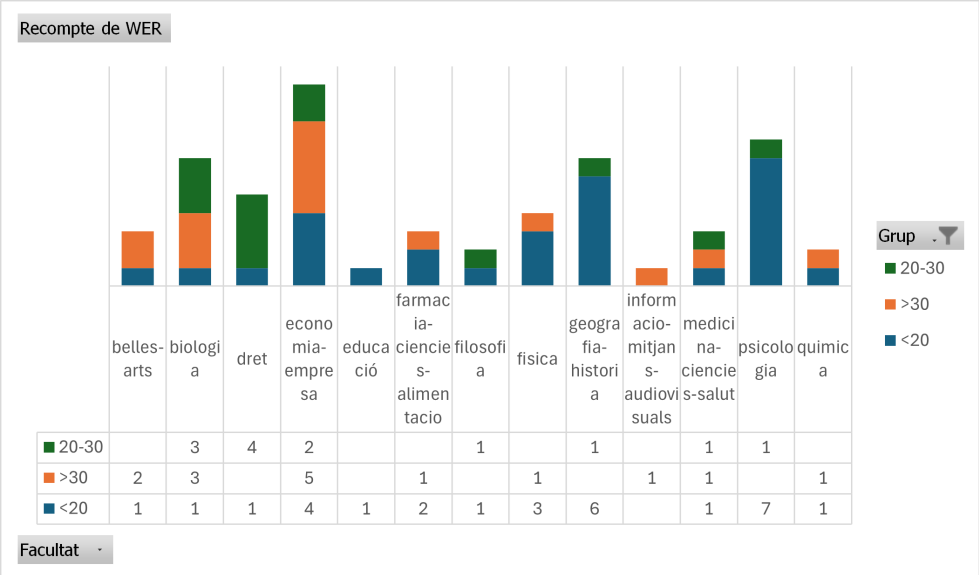
El grau de Química i el màster de Seguretat alimentària continuen sent els graus amb més error, tant en caràcters com en paraules. El grau de Ciències ambientals també es troba entre els que tenen més errors de la segona categoria, en ambdues mètriques. Pel que fa al grup amb menys errors, és esperable que els que no han presentat cap caràcter erroni tampoc hagin tingut cap paraula equivocada. El que es pot destacar d'aquest últim grup és que el següent grau amb més errors ja supera l'1 %, mentre que en la mètrica CER es constatava més precisió, amb uns quants graus i màsters amb uns resultats inferiors a l'1 %.

Fer aquestes mètriques ha permès fer una primera aproximació a la detecció de les facultats que fan servir un vocabulari específic més desconegut per al model base de l'SCRIBAL i que, per tant, cal entrenar amb noves dades. Si bé les mètriques CER indiquen una bona transcripció en general, les mètriques WER mostren que hi ha 15

[9]Es pot veure la taula completa a l'Apèndix A, taula 8.

graus i màsters que requereixen més entrenament de dades.

Si es traslladen aquestes dades en una taula dinàmica, es pot fer una agrupació de quines facultats han obtingut més errors als seus graus. Els resultats, presents a la figura 1, mostren que les facultats d'Economia i empresa, Biologia i Belles arts, tenen més graus en el rang > 30 % de WER —5, 3 i 2 respectivament.



	belles-arts	biologia	dret	economia-empresa	educació	farmacia-ciencies-alimentacio	filosofia	fisica	geografia-historia	informacio-mitjans-audiovisuals	medicina-ciencies-salut	psicologia	quimica
20-30		3	4	2			1		1		1	1	
>30	2	3		5		1		1		1	1		1
<20	1	1	1	4	1	2	1	3	6		1	7	1

Figura 1: Resultats desglossats per facultats segons els resultats WER.

Tot i que hi ha altres facultats que tenen un grau o màster en la franja de *transcripció pobra* —com les de Farmàcia i ciències de l'alimentació, Física, Informació i mitjans audiovisuals, Medicina i ciències de la salut i Química—, aquestes dades ofereixen un punt de partida per a començar a millorar de forma escalada el model en tots els àmbits universitaris possibles.

Cal afegir que hi ha 29 graus, una mica més de la meitat, amb un error inferior al 20 %, és a dir, que s'han transcrit correctament.

4.1.2 Anàlisi qualitativa

Malgrat tot, els resultats de l'anàlisi quantitativa inicial es poden deure a diferents factors que no són únicament errors en el processament terminològic del model. És per això que cal veure quins són els errors que s'han produït en la transcripció, quins són els més comuns i, sobretot, veure realment quina terminologia específica és la que més manca en el model SCRIBAL per poder-lo fer funcionar en tots els àmbits universitaris.

Com que la facultat d'Economia i empresa és la que ha tingut uns resultats WER més elevats, s'ha començat per aquí a analitzar qualitativament la transcripció. Les puntuacions superiors a 30 % s'han donat a tres graus —ADE, Empresa internacional, Economia (tg1019)— i dos màsters —Ciències actuarials financeres (md503) i Internacionalització. A continuació es mostra una taula (3) recollint els seus errors terminològics.

Grau ADE	
document d'Excel	document d'**èxtil**
Màrqueting Estratègic	**Marketing** estratègic
Màrqueting internacional	**Marketing** internacional
Grau en Empresa Internacional	
Sostenibilitat Global	**Sustainabilitat** global
Grau en Economia (tg1019)	
Cap error.	
Màster en ciències actuarials financeres (md503)	
Dret Fiscal, Bancari i Borsari	dret fiscal bancari i **bursari**
Operacions Actuarials sobre Grups i Invalidesa	operacions actuarials sobre grups **i assegurances,** invalidesa
Màster en internacionalització	
Dret, Economia i Globalització	dret **a** economia i globalització

Taula 3: Taula dels graus amb una puntuació < 30 % i els errors qualitatius.

L'anàlisi qualitativa mostra que hi ha relativament pocs errors terminològics. En aquests graus només s'han transcrit cinc mots de forma diferent, dos dels quals són el mateix —*màrqueting* escrit en anglès. A més, als màsters la transcripció ha generat dues confabulacions, una a cada un. Però el que sorprèn més és que el grau en Economia s'ha transcrit igual que l'original.

La conclusió a què s'arriba fent un balanç general dels errors és que n'hi ha d'altres que no són específicament terminològics i que no s'han tingut en compte en el moment de fer la normalització dels textos prèvia a l'anàlisi quantitativa. Per exemple, el grau en Economia no té cap error terminològic, però la majoria de les assignatures tenen

un número romà al títol. El model ha fet la transcripció àrab d'aquests números —perquè oralment no es pot distingir una forma de l'altra—, cosa que ha fet augmentar el percentatge d'error.

Aquest tipus d'errors també s'han esdevingut en el grau de Química i en el màster de Seguretat Alimentària, ambdós amb el CER i WER més elevat de tot el corpus. En el primer, tot i haver-hi només tres errors terminològics, hi ha hagut una confabulació molt llarga al final del text; en el segon, la transcripció ha obviat vuit línies de text. Aquestes interferències han fet augmentar la taxa d'error.

Gràcies a l'anàlisi qualitativa s'ha pogut veure quin tipus d'errors han estat els més freqüents i s'han classificat per tipus. En total s'han trobat sis errors de transcripció:

- **Números romans (N):** Moltes assignatures es divideixen en parts que es fan en diferents cursos. Això s'indica a través de números romans. El transcriptor no detecta quan són números romans o quan són números àrabs i sempre transcriu els segons.

  (1) Gold-Standard: Història de l'Art III
  Transcripció: Història de l'Art **3**

- **Puntuació (P):** Si bé a través de l'entonació i la durada de les pauses la transcripció de punts i comes és prou acceptable, falla algunes vegades. Sobretot no és capaç de diferenciar entre els anteriors i dos punts, guions o parèntesis.

  (2) Gold-Standard: Biosensors i Lab-on-a-chip
  Transcripció: Biosensors i **Lab on a Chip**

- **Confabulacions (C):** Un dels errors més freqüents que s'han detectat ha estat l'aparició de confabulacions —paraules o frases que no s'han dit a l'enregistrament. Tot i que n'han aparegut al mig d'algunes frases, com a l'exemple (3) del grau de Restauració i conservació de béns culturals, la majoria han tingut lloc al final de la transcripció, en moments de silenci (exemple (4) del màster d'Intervenció psicosocial).

(3) Gold-Standard: Restauració de pintura sobre fusta.
Transcripció: Restauració de **pintura sobre tela.** Pintura sobre fusta.

(4) Gràcies! Gràcies! Gràcies!

- **Substitucions (S):** En alguns casos, hi ha hagut paraules que no s'han transcrit erròniament, sinó que s'han canviat per una altra completament diferent.

(5) Gold-Standard: Història de l'Art II: Art Medieval, Renaixentista i Barroc
Transcripció: Art medieval, renaixentista i **antiga**.

- **Omissions (O):** Un altre error detectat han estat les omissions. El transcriptor no ha detectat algunes parts dels enregistraments i no les ha transcrit, cosa que ha fet augmentar el còmput d'elisions. A l'hora de recomptar-les, s'ha fet la suma de línies totals omeses.

- **Errors terminològics (T):** I finalment, també hi ha hagut errors terminològics, en què els mots transcrits tenien algun error ortogràfic o eren formalment similars al Gold-Standard. S'analitzaran en profunditat al següent apartat.

Fora d'aquesta classificació es podria afegir un apartat d'*altres*, ja que s'han detectat tres fenòmens que no es poden considerar errors, però que val la pensa comentar. El primer són les **sigles i els acrònims**. En aquest corpus s'han lletrejat tots lletra a lletra i s'han transcrit, majoritàriament, de forma correcta. Ara bé, pot ser que algunes es pronunciïn com acrònims, cosa que potser provoca algun error a la transcripció.

El segon són els **símbols matemàtics**. En algunes assignatures els títols tenen el símbol +, que s'ha transcrit amb la paraula *més*. No es pot considerar un error, però s'ha de tenir en compte si apareix sovint en graus tècnics.

L'últim fenomen a tenir en compte són les **correccions**. Hi ha hagut algun cas en què el corpus de referència tenia errors ortogràfics que la transcripció ha corregit. Això no es pot tenir en compte amb les mètriques d'avaluació.

Tant els errors de números romans com els de puntuació es poden corregir a través de la normalització del text per fer càlculs WER i CER més acurats. És per això que no

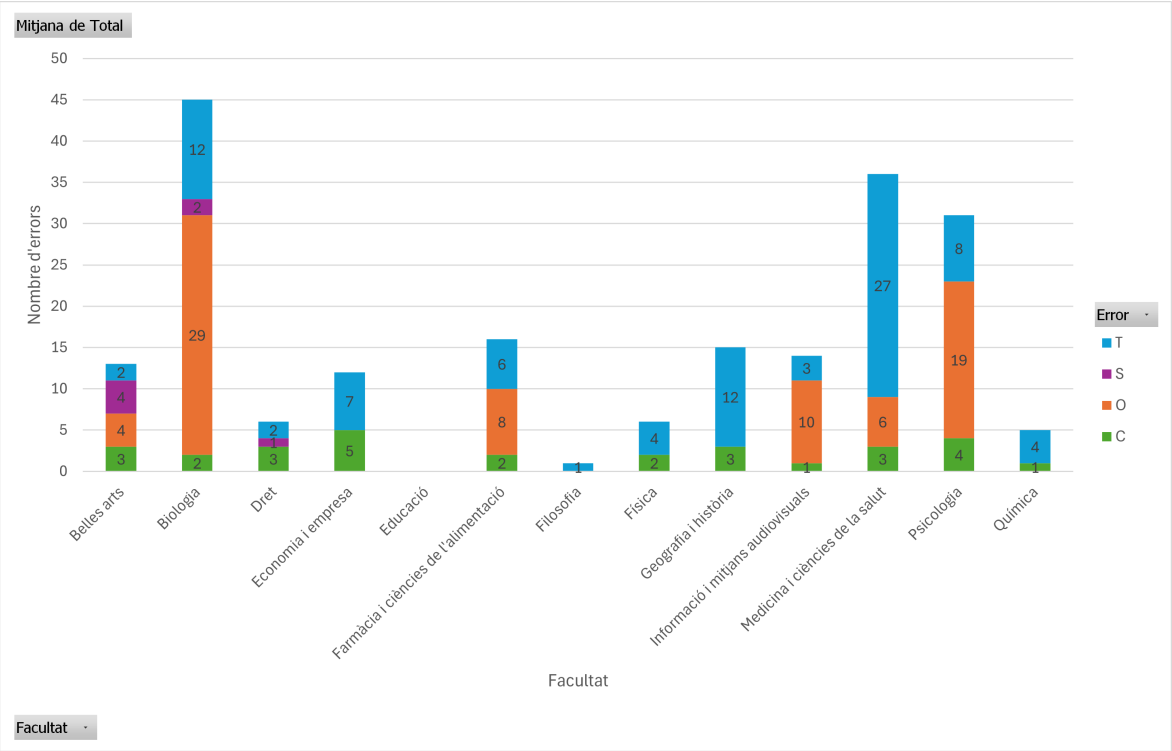


Figura 2: Resum dels errors de cada facultat, desglossats per tipus (T: terminològics, S: substitucions, O: omissions, C: confabulacions).

s'han inclòs en el recompte comparatiu d'errors més freqüents [10], mostrat a l'Apèndix B, figura 4.

El recompte total d'errors mostra que els errors terminològics són els més freqüents (88), seguits de les omissions (76) i de les confabulacions (29). Les substitucions són les menys freqüents (7). Si es mira quines facultats han obtingut més errors (Annex B, figura 5), es veu que Biologia és la facultat que acumula més errors en total. Tanmateix, si es desglossen aquests resultats, es veu que la facultat que acumula més errors terminològics és la de Medicina (27), amb el grau d'Odontologia i els màsters de Recerca clínica i Competències mèdiques avançades. Les facultats d'Educació i Filosofia tenien poques dades, cosa que explica que no tinguin gaires errors. La figura 2 resumeix aquestes dades[11].

[10] Les omissions es poden intentar evitar a través d'un filtre, però no és un mètode que en garanteixi la correcció. A més, han tingut un impacte molt gran en el còmput d'errors

[11] Si es volen veure els resultats totals per a cada error i per a cada facultat, vegeu B, taules 6, 7, 8 i 9.

4.1.3 Anàlisi dels errors terminològics

Els errors terminològics que s'han detectat en el corpus són molt variats:

- S'ha notat certa confusió amb les preposicions *en* i *amb*, segurament a causa de la seva similitud fonètica en català: [ən] i [əm].

  (6) GS: amb Perspectiva de Gènere
  TR: **en** perspectiva de gènere

- Algunes paraules es transcriuen en un altre idioma, tant pot ser castellà (7) o en anglès (8).

  (7) GS: metabolisme
  TR: **metabolismo**

  (8) GS: màrqueting
  TR: **marketing**

- Les vocals que pateixen reducció vocàlica a vegades no es transcriuen correctament.

  (9) GS: forense
  TR: foren**ça**[12]

- Algunes consonants es confonen.

  (10) GS: mostreig
  TR: mostrei**x**

  (11) GS: joiers
  TR: jo**ll**ers

[12] Aquest cas és curiós perquè ha adaptat la grafia de la consonant anterior al so contingu.

- Pot tenir problemes amb els accents i les dièresis.

(12) GS: mèdicament
TR: m**e**dicament

(13) GS: aqüicultura
TR: aq**u**icultura

- Hi ha paraules que desconeix.

(14) GS: Sjörgen
TR: **Sioracren**

Amb l'entrenament específic de terminologia s'espera resoldre aquests errors i millorar els resultats.

4.2 Avaluació específica

4.2.1 Anàlisi quantitativa

Un cop readaptades les mètriques WER i CER i aplicades al Corpus Medicina, els resultats obtinguts han estat més acurats. Es poden veure desglossats a la taula 4.

Corpus	**WER**	**CER**
UBterm	142,86	12,98
PD	57,16	34,35
Assignatures	21,04	3,52

Taula 4: Resultats WER i CER de l'anàlisi específica de medicina.

A diferència de la primera anàlisi, es pot veure que les millores en el codi han permès rebaixar les taxes d'error, aproximant-lo cada cop més a un error estrictament terminològic i no a causa d'altres factors.

Si recordem els tipus de dades que s'extreuen de cada corpus:

- L'*UBterm* és una llista de paraules de vocabulari específic sense context.

- Els *PD* són documents amb paraules específiques en el seu context.
- Les *Assignatures* es troben entre els dos anteriors: són segments de vocabulari específic en un context més reduït i amb elements gramaticals que els relacionen.

no sembla que hi hagi una relació específica entre context i millor actuació. Es podria pensar que les paraules descontextualitzades (UBterm: 142,86) obtenen pitjors resultats, però, la gradació que s'esperaria en l'error de la resta —les assignatures en tindrien més que els PD—, no s'esdevé. De fet, els PD (WER $>$ 50 %) obtenen pitjors resultats que les assignatures (WER $<$ 30 %).

4.2.2 Anàlisi qualitativa

Gràcies a les modificacions que s'han fet a la normalització del text, s'han aconseguit reduir els errors no terminològics a gairebé zero. No hi ha hagut omissions, però sí que hi ha hagut confabulacions en les Assignatures: al final d'1 i 2 han aparegut les paraules "cuts, actualització dalej." i "Gràcies!" respectivament. També n'hi ha hagut una a l'UBterm: "Fiots, Esportes, Osteogenes, Ot Spain, Oncogenic, Oncogen, Oncogen, Oncogen, Oncogen, Oncogen, Oncogen, Oncoins, Oncogen, Oncogen, Oncogen, Oncogen, Oncogen, Oncogen, Onc Pixo, AMCLI, Osteorganizaciones, I nesse video."

Pel que fa als errors terminològics, n'hi ha hagut de tots tipus: d'accentuació (15), de vocals (16), de consonants (17) i de desconeixement de la paraula (18).

(15) GS: Conèixer
TR: Con**e**ixer

(16) GS: butllofa
TR: b**o**tllofa

(17) GS: caquèxia
TR: caque**cc**ió

(18) GS: schwanoma
TR: **xu**anoma

4.3 Entrenament del model

Els resultats de l'entrenament han mostrat millores importants en el model pel que fa al vocabulari específic. Tant la WER com la CER han disminuït considerablement, cosa que indica que les transcripcions han estat molt més acurades i el model ha adquirit la terminologia del grau de medicina.

Com es pot veure a les taules 5 i 6, les taxes d'error de les paraules contextualitzades —els plans docents i les assignatures— es troben al voltant de l'1,15 per la WER i de 0,40 per la CER. El vocabulari descontextualitzat, però, continua obtenint una WER molt alta (> 30 %), malgrat que s'hagi rebaixat un 72,21 %.

Corpus	**WER inicial**	**WER resultant**
UBterm	142,86	39,70
PD	57,16	1,00
Assignatures	21,04	1,30

Taula 5: Resultats WER del Corpus Medicina desglossat abans i després de l'entrenament de vocabulari específic.

Corpus	**CER inical**	**CER resultant**
UBterm	12,98	11,30
PD	34,35	0,30
Assignatures	3,52	0,50

Taula 6: Resultats CER del Corpus Medicina desglossat abans i després de l'entrenament de vocabulari específic.

Els errors que ha comès, tot i que amb menys freqüència, continuen sent els que s'havien detectat fins ara: confusió de vocals (19), confusió de consonants (20) i manca d'accents (21), principalment.

(19) GS: acalàcia
TR: **e**calàcia

(20) GS: limfòcit
TR: li**n**fòcit

(21) GS: alvèol
TR: alv**e**ol

Altrament, el model ha empitjorat la seva actuació general, tot i la millora en la terminologia mèdica. Els resultats obtinguts de la transcripció d'un corpus general[13] —amb una WER de 9,7 i una CER de 5,7— empitjoren després d'entrenar el model amb la terminologia específica. Aquest corpus concret ha obtingut una WER de 10,9 i una CER de 6,2. Tot i que el canvi és lleu, és quelcom a tenir en compte si es volen incrementar els entrenaments per a millorar el vocabulari amb més dades perquè no es vagi acumulant l'error i acabi empitjorant la seva actuació en àmbits on anteriorment no presentava problemes.

4.4 Agrupació per branques de coneixement

Els grups resultants de l'agrupació per branques de coneixement han estat prou acurats amb la divisió per facultats que es troba a les universitats: informàtica, geologia, biologia, ciències humanes, economia i dret, educació, ciències exactes...

Tanmateix, han sorgit alguns aparellaments que no sembla que tinguin relació. Els grups 5-7 i 9-12 relacionen graus de branques molt diferents, com per exemple Educació musical i Palinologia, una disciplina de la botànica.

Per presentar de forma visual les relacions entre els grups, es va crear un dendrograma, reproduït a la figura 3.

Aquest arbre classifica els graus en 12 grups que tenen el vocabulari més pròxim. El grup vermell més extens mostra que els diferents grups de paraules estan molt allunyats entre ells (entre 125 i 150). És per això que dins d'aquesta categoria hi ha Informàtica, Ecologia i Educació musical. Al grup marró hi passa alguna cosa similar: s'hi troben assignatures com Geologia, Podologia o Llibre manuscrit.

[13] Aquest corpus s'ha creat des de l'SCRIBAL per al seu desenvolupament i millora, de manera que no es pot trobar publicat.

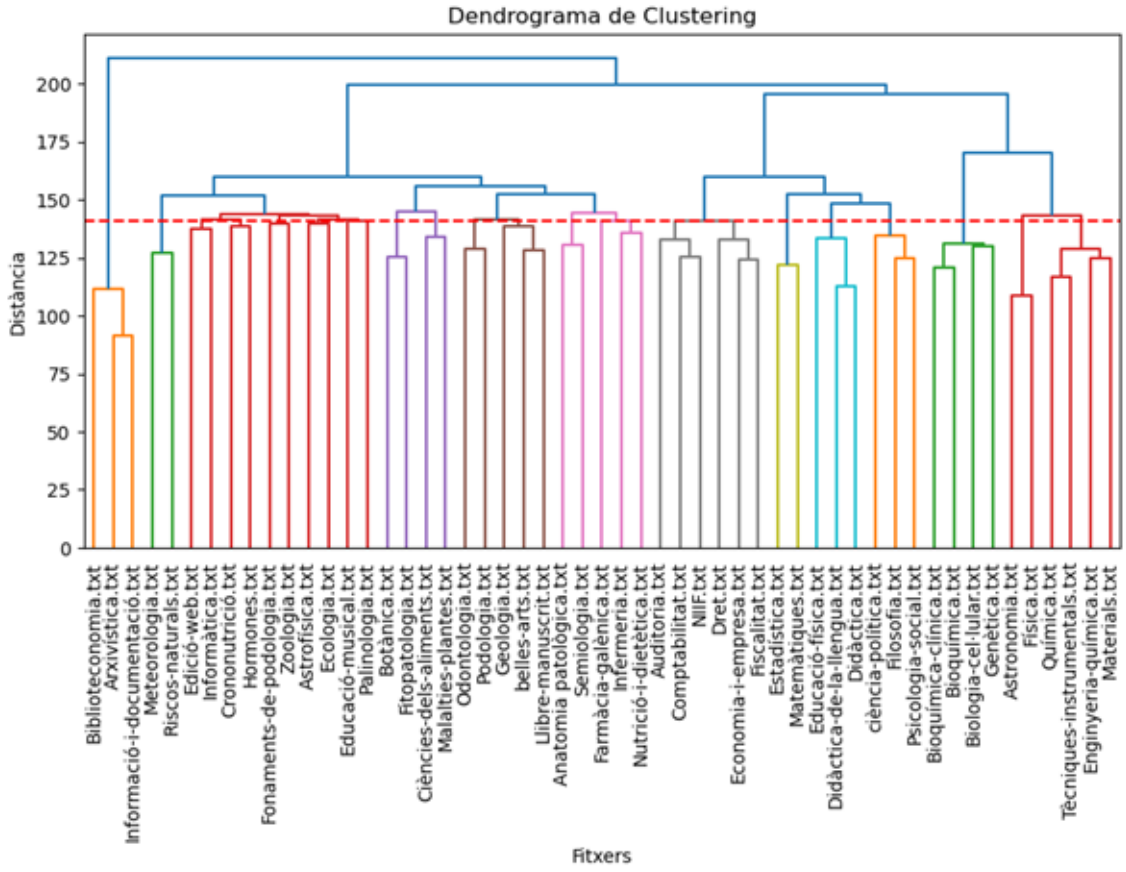


Figura 3: Dendrograma de les assignatures amb vocabulari comú.

Aquest tipus d'agrupació també s'ha esdevingut en el grup lila clar, però en menys mesura, en què entre diferents graus de l'àmbit de la salut s'hi troba Semiologia. Dins dels grups lila fosc i blau hi ha un valor que també despunta: d'una banda, en l'àmbit de la botànica apareix el grau de Ciències dels aliments; d'altra banda, l'educació física s'inclou en l'àmbit de la didàctica.

En canvi, els grups que han resultat amb un vocabulari més pròxim (al voltant de 125 o inferior), s'han agrupat de forma similar a la de les facultats universitàries:

- **Taronja 1**: Biblioteconomia, Arxivística, Informació i documentació.
- **Verd fosc 1**: Meteorologia, Riscos naturals.
- **Lila fosc**: Botànica, Fitopatologia, Ciències dels aliments, Malalties de les plantes.
- **Gris**: Comptabilitat, NIIF, Dret, Economia i empresa.
- **Verd clar**: Estadística, Matemàtiques.

- **Blau**: Educació física, Didàctica de la llengua, Didàctica.
- **Taronja 2**: Ciència política, Filosofia, Psicologia social.
- **Verd fosc 2**: Bioquímica clínica, Bioquímica, Biologia, cel·lular, Genètica.
- **Vermell 2**: Astronomia, Física, Química, Tècniques instrumentals, Enginyeria química, Materials.

Aquesta classificació mostra que hi ha molts graus que comparteixen vocabulari específic. En general, els graus de ciències acostumen a estar molt relacionats entre si. Això porta a pensar que una adaptació del vocabulari per a facultats —per exemple, física, química i enginyeria— seria més adequada i òptima que no pas una adaptació específica per a cada grau, cosa que suposaria més feina i recursos a l'hora de desenvolupar l'adaptació terminològica.

Tot i això, no s'ha de perdre de vista que els graus científics estan molt més representats que els graus d'humanitats i arts. També és per això que la classificació ha estat més exacta en el primer àmbit, mentre que les segones s'han trobat agrupades en un calaix de sastre amb altres disciplines més o menys properes. Amb una representació més justa de cada àrea o amb més dades de les infrarepresentades, es podria veure de forma més exacta la validesa d'adaptar terminològicament les àrees de coneixement segons les facultats de la universitat.

Per últim, cal tenir en compte que hi ha alguns ensenyaments que pertanyen a una facultat concreta, però que el seu contingut es podria classificar com a propi d'una altra. Per exemple, el màster de Lògica pura aplicada es fa a la facultat de Filosofia. Ara bé, si es mira el seu vocabulari, és més probable que encaixi millor amb el de la facultat de Matemàtiques. És per això que podria ser interessant oferir a cada ensenyament, independentment de la facultat a la qual pertany, l'opció de vincular-se al model de vocabulari específic que més s'adigui a les seves necessitats.

5 CONCLUSIONS

El model d'SCRIBAL, tot i ser un model que necessita millores en l'àmbit terminològic de les especialitats universitàries, té una actuació relativament bona: més de la meitat dels ensenyaments de la Universitat de Barcelona es poden transcriure amb un resultat més que acceptable.

Tanmateix, els errors de vocabulari específic no són els únics que cal corregir. Si bé són els més freqüents, un 44 % dels errors totals, la suma d'omissions i confabulacions suposa el 52,5 %, fet que s'ha de tenir en compte a l'hora de refinar el model. Més enllà de la necessària adaptació terminològica, cal intentar reduir aquests errors que a vegades són més notoris en una transcripció que no pas els ortogràfics.

Pel que fa a l'adaptació terminològica, s'ha vist clarament com l'entrenament amb les dades de la facultat de Medicina ha permès reduir considerablement les taxes d'error, sobretot quan el vocabulari es troba contextualitzat. El model continua presentant alguns problemes quan el vocabulari no està contextualitzat, especialment si són paraules molt tècniques. A més, malgrat la millora en el vocabulari tècnic, el model ha empitjorat lleugerament en la seva actuació en corpus generals.

L'entrenament amb dades sintètiques ha resultat ser un èxit en aquest treball, tenint en compte que l'avaluació resultant de l'entrenament amb TTS s'ha fet amb dades reals i les transcripcions han obtingut unes taxes WER i CER molt més baixes que les obtingudes abans de l'entrenament. Això pot resultar molt útil en futures adaptacions, ja que els recursos, sobretot de temps i costos, es podran optimitzar millor fent servir sistemes TTS per fer *fine-tuning*. Tot i això, cal ser caut davant l'automatització total d'aquests processos: no s'ha de perdre de vista que ha de ser capaç de reconèixer tota mena de veus humanes en un entorn sorollós; l'entrenament no es pot basar únicament en dades perfectes.

Finalment, s'ha vist que pot ser més útil i òptim adaptar la terminologia segons la branca de coneixement a la qual pertany. La majoria d'ensenyaments universitaris que es troben a la mateixa facultat comparteixen bona part del vocabulari, de manera que es podria mirar de crear i entrenar models diferents per a cada una d'elles. Així, cada grau o màster —i fins i tot assignatura— podria escollir aquell model que proporciona

un millor resultat per al tipus de contingut que ofereix.

5.1 Futures línies de recerca

Després d'una recerca com la que ha suposat aquest treball, s'obren noves línies per on continuar investigant i desenvolupant el model de transcripció de l'SCRIBAL.

El més immediat seria provar el model entrenat en una aula en directe i veure la seva actuació real. Les dades que proporcionaria fer una prova pilot l'ajudarien a millorar encara més.

Una altra possible línia d'investigació seria detectar els límits —si n'hi ha— de l'entrenament amb síntesi de la parla. Fer una comparació de l'actuació de models entrenats únicament amb TTS i d'altres, únicament amb parla real, permetria valorar els beneficis i inconvenients de fer servir tecnologies de la parla en tots els processos d'entrenament d'un model.

Finalment, es podria intentar trobar la manera més eficient de millorar l'actuació en la terminologia específica sense perjudicar l'actuació general. Juntament amb la idea d'agrupar els ensenyaments amb diferents models entrenats segons el seu vocabulari, es podria aplicar la proposta de Lall and Liu (2024) d'esbiaixar contextualment aquests models per tal d'obtenir una transcripció més precisa amb menys costos d'entrenament.

REFERÈNCIES

Barcelona Supercomputing Center, T. L. T. U. f. (2024). Natural and efficient tts in catalan: using matcha-tts with the catalan language.

Bell, P., Yamamoto, H., Swietojanski, P., Wu, Y., McInnes, F., Hori, C., and Renals, S. (2013). A lecture transcription system combining neural network acoustic and language models. In *In Proc. Interspeech 2013*.

Eric-Urban (2025). Test accuracy of a custom speech model - Speech service - Azure AI services. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-custom-speech-evaluate-data?pivots=ai-foundry-portal.

Fügen, C., Waibel, A., and Kolss, M. (2007). Simultaneous translation of lectures and speeches. *Machine Translation*, 21:209–252.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Huh, J., Park, S., Lee, J. E., and Ye, J. C. (2023). Improving medical speech-to-text accuracy with vision-language pre-training model. https://arxiv.org/abs/2303.00091.

Joshi, R. and Singh, A. (2022). A simple baseline for domain adaptation in end to end ASR systems using synthetic data. In Malmasi, S., Rokhlenko, O., Ueffing, N., Guy, I., Agichtein, E., and Kallumadi, S., editors, *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 244–249, Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/2022.ecnlp-1.28/.

Jurafsky, D. and Martin, J. H. (2025a). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, chapter Automatic Speech Recognition and Text-to-Speech. draft, 3rd online manuscript released january 12 edition. https://web.stanford.edu/~jurafsky/slp3/16.pdf.

Jurafsky, D. and Martin, J. H. (2025b). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, chapter Large Language Models. draft, 3rd online manuscript released january 12 edition. https://web.stanford.edu/~jurafsky/slp3/10.pdf.

Kulkarni, R. V., Shilimkar, A., Bandi, S., Bhegade, S., and Dadmal, J. (2024). Transcription, translation and summarization to improve educational understanding. In *2024 IEEE 4th International Conference on ICT in Business Industry and Government (ICTBIG)*, pages 1–7.

Lall, V. and Liu, Y. (2024). Contextual biasing to improve domain-specific custom vocabulary audio transcription without explicit fine-tuning of whisper model. https://arxiv.org/abs/2410.18363.

Lamel, L., Adda, G., Bilinski, E., and Gauvain, J.-L. (2005). Transcribing lectures and seminars. In *Interspeech 2005*, pages 1657–1660.

Llisterri, J. (2009). Les tecnologies de la parla. *Llengua, societat i comunicació*, pages 11–19. https://revistes.ub.edu/index.php/LSC/article/view/3319.

Maulik, U. B., Mitra, P., and Sarkar, S. (2024). Enhancing Domain-Specific ASR Performance Using Finetuning and Zero-Shot Prompting: A Study in the Medical Domain. *The 2024 Sixth Doctoral Symposium on Intelligence Enabled Research (DoSIER 2024)*, 3900:172–180. https://ceur-ws.org/Vol-3900/Paper14.pdf.

Mehta, S., Tu, R., Beskow, J., Éva Székely, and Henter, G. E. (2024). Matcha-tts: A fast tts architecture with conditional flow matching.

Müller, M., Nguyen, T. S., Niehues, J., Cho, E., Krüger, B., Ha, T.-L., Kilgour, K., Sperber, M., Mediani, M., Stüker, S., and Waibel, A. (2016). Lecture translator - speech translation framework for simultaneous lecture translation. In DeNero, J., Finlayson, M., and Reddy, S., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, San Diego, California. Association for Computational Linguistics.

Murphy, J. and Hallinger, P. (1985). Transcript analysis: A tool for improving quality and equity in high school programs. *The High School Journal*, 69(2):132–138.

OpenAI (2022). whisper. *GitHub repository*. https://github.com/openai/whisper.

Pawlowicz, D., Weber, J., and Dukino, C. (2025). Effectiveness of whisper's fine-tuning for domain-specific use cases in the industry. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, pages 1396–1403. INSTICC, SciTePress.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. https://arxiv.org/abs/2212.04356.

Softcatalà (2023). whisper-ctranslate2. *GitHub repository*. https://github.com/Softcatala/whisper-ctranslate2/tree/main.

Tran, M., Pang, Y., Paul, D., Pandey, L., Jiang, K., Guo, J., Li, K., Zhang, S., Zhang, X., and Lei, X. (2025). A domain adaptation framework for speech recognition systems with only synthetic data. https://arxiv.org/abs/2501.12501.

Vásquez-Correa, J. C., Arzelus, H., Martin-Doñas, J. M., Arellano, J., Gonzalez-Docasal, A., and Álvarez, A. (2023). When whisper meets tts: Domain adaptation using only synthetic speech data. In Ekštein, K., Pártl, F., and Konopík, M., editors, *Text, Speech, and Dialogue*, pages 226–238. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40498-6_20.

Apèndix A:

Resultats CER i WER de l'avaluació general

% d'error	CER	Grau/Màster
> 30 %	48,61	seguretat-alimentària
> 30 %	45,53	química
> 30 %	33,70	ecologia-gestió-restauració
20-30 %	27,75	gestió-informació-documentació-digital
20-30 %	24,98	ciències-ambientals
20-30 %	23,55	estadística
20-30 %	18,68	intervenció-psicosocial
20-30 %	16,04	conservació-restauració-béns-culturals-tg1118
20-30 %	14,82	bioquímica
20-30 %	12,09	energies-renovables-sostenibilitat-energètica
20-30 %	10,55	producció-recerca-artística
20-30 %	10,33	conservació-restauració-béns-culturals-tg1003
20-30 %	9,88	antropologia-social-cultural
20-30 %	9,68	recerca-clínica
20-30 %	9,49	nutrició-humana-dietètica
20-30 %	8,96	psicologia
20-30 %	7,48	ade
20-30 %	7,43	geografia
20-30 %	6,62	competències-mèdiques-avançades
20-30 %	5,9	biotecnologia
20-30 %	5,80	biologia
20-30 %	5,74	empresa-internacional
20-30 %	5,38	sociologia
20-30 %	5,36	odontologia
20-30 %	5,30	ciències-actuarials-financeres-md503
20-30 %	5,26	economia-tg1019
20-30 %	5,09	direcció-estratègica-seguretat
20-30 %	4,69	història-art
20-30 %	4,62	nanociència-nanotecnologia
20-30 %	4,39	dret

% d'error	CER	Grau/Màster
20-30 %	4,36	farmàcia
< 20 %	4,35	gestió-administració-pública
< 20 %	4,28	ciutadania-drets-humans-ètica-política
< 20 %	4,03	enginyeria-biomèdica
< 20 %	3,85	seguretat-salut-prevenció-riscos-laborals
< 20 %	3,76	internacionalització
< 20 %	3,70	recerca-comportament-cognició
< 20 %	3,59	criminologia-política-criminal
< 20 %	3,54	història
< 20 %	3,37	dret-empresa-negocis
< 20 %	3,25	biodiversitat
< 20 %	3,17	psicologia-general-sanitària
< 20 %	3,07	economia-regulació-competència
< 20 %	2,61	gestió-desenvolupament-persones-equips
< 20 %	2,50	economia-tg1115
< 20 %	2,33	direcció-empreses-esport
< 20 %	2,24	ciències-actuarials-financeres-md5dp
< 20 %	1,82	arqueologia
< 20 %	1,42	meteorologia
< 20 %	1,42	química-orgànica
< 20 %	0,61	estudis-llatinoamericans
< 20 %	0,56	psicogerontologia
< 20 %	0,56	psicologia-educació-mipe
< 20 %	0,52	microbiologia-avançada
< 20 %	0,35	lògica-pura-aplicada
< 20 %	0,00	direcció-gestió-centres-educatius
< 20 %	0,00	geografia-canvi-global

Taula 7: Resultats CER de l'anàlisi general, segons el seu percentatge d'error.

% d'error	WER	Grau/Màster
$>$ 30 %	67,72	química
$>$ 30 %	47,44	seguretat-alimentària
$>$ 30 %	46,86	conservació-restauració-béns-culturals-tg1108
$>$ 30 %	45,69	biologia
$>$ 30 %	39,04	ecologia-gestió-restauració
$>$ 30 %	37,78	biotecnologia
$>$ 30 %	34,62	conservació-restauració-béns-culturals-tg1003
$>$ 30 %	33,55	ade
$>$ 30 %	32,39	empresa-internacional
$>$ 30 %	32,22	ciències-actuarials-financeres-md503
$>$ 30 %	32,22	gestió-informació-documentació-digital
$>$ 30 %	32,04	odontologia
$>$ 30 %	31,03	economia-tg1019
$>$ 30 %	30,86	nanociència-nanotecnologia
$>$ 30 %	30,77	internacionalització
20-30 %	29,41	sociologia
20-30 %	29,17	direcció-estratègica-seguretat
20-30 %	27,52	ciències-ambientals
20-30 %	27,03	història-art
20-30 %	26,74	dret-empresa-negocis
20-30 %	26,00	ciutadania-drets-humans-ètica-política
20-30 %	25,00	dret
20-30 %	24,56	bioquímica
20-30 %	24,18	recerca-clínica
20-30 %	23,81	criminologia-política-criminal
20-30 %	22,86	biodiversitat
20-30 %	20,95	intervenció-psicosocial
20-30 %	20,00	economia-regulació-competència
$<$ 20 %	19,67	direcció-empreses-esport
$<$ 20 %	19,05	gestió-desenvolupament-persones-equips
$<$ 20 %	18,94	producció-recerca-artística
$<$ 20 %	17,65	seguretat-salut-prevenció-riscos-laborals
$<$ 20 %	17,01	geografia
$<$ 20 %	16,88	energies-renovables-sostenibilitat-energètica

% d'error	WER	Grau/Màster
< 20 %	14,08	farmàcia
< 20 %	14,06	economia-tg1115
< 20 %	13,87	antropologia-social-cultural
< 20 %	12,56	competències-mèdiques-avançades
< 20 %	12,14	nutrició-humana-dietètica
< 20 %	11,14	psicologia
< 20 %	9,57	gestió-administració-pública
< 20 %	9,33	arqueologia
< 20 %	9,09	meteorologia
< 20 %	7,89	ciències-actuarials-financeres-md5dp
< 20 %	7,65	història
< 20 %	7,62	recerca-comportament-cognició
< 20 %	6,92	enginyeria-biomèdica
< 20 %	6,25	química-orgànica
< 20 %	5,75	estudis-llatinoamericans
< 20 %	4,00	psicologia-general-sanitària
< 20 %	3,73	estadística
< 20 %	3,72	psicologia-educació-mipe
< 20 %	2,70	lògica-pura-aplicada
< 20 %	1,64	microbiologia-avançada
< 20 %	1,61	psicogerontologia
< 20 %	0,00	direcció-gestió-centres-educatius
< 20 %	0,00	geografia-canvi-global

Taula 8: Resultats WER de l'anàlisi general, segons el seu percentatge d'error.

Apèndix B:

Gràfics resultants de l'avaluació qualitativa

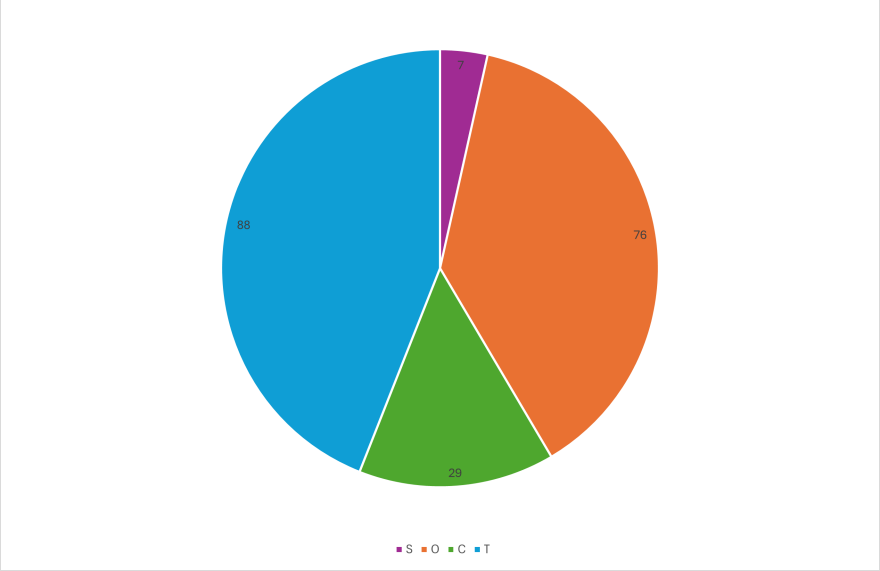


Figura 4: Errors totals segons el tipus (T: terminològics, S: substitucions, O: omissions, C: confabulacions).

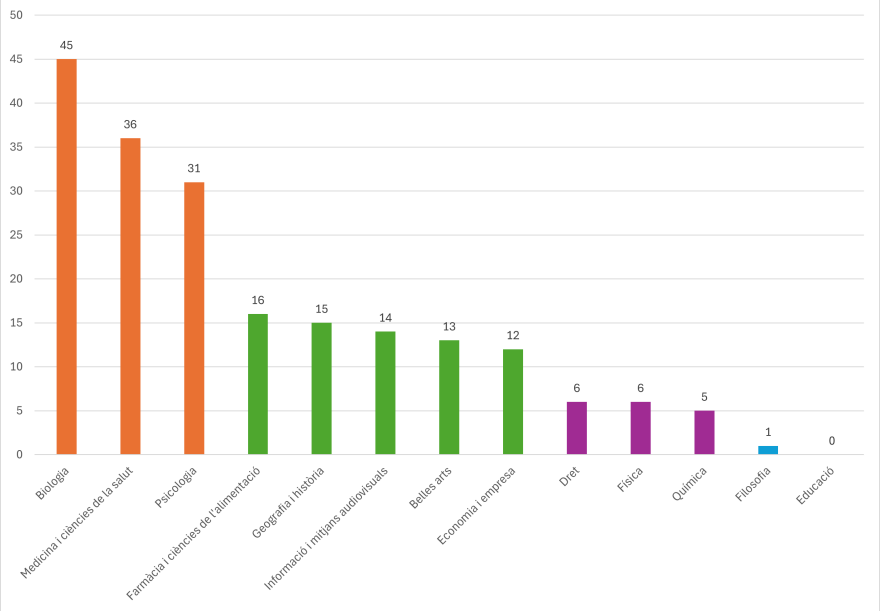


Figura 5: Quantitat d'errors total (y) segons la facultat (x).

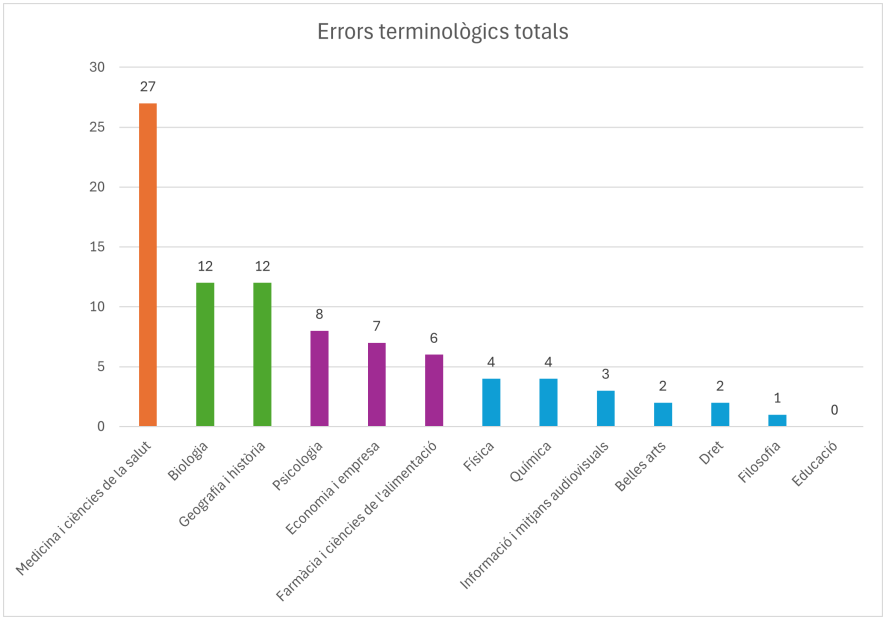


Figura 6: Quantitat d'errors terminològics totals (y) segons la facultat (x).

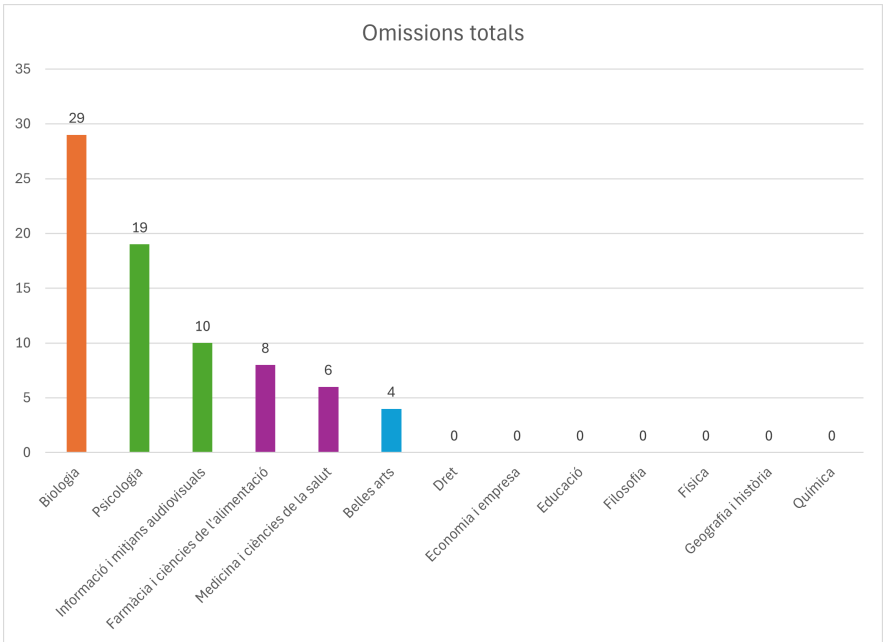


Figura 7: Quantitat d'omissions totals (y) segons la facultat (x).

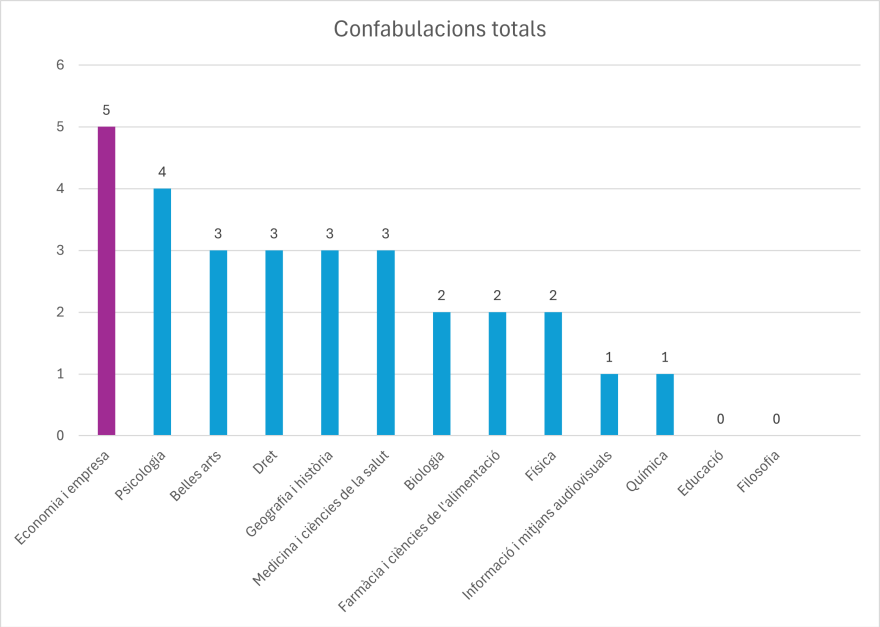


Figura 8: Quantitat de confabulacions totals (y) segons la facultat (x).

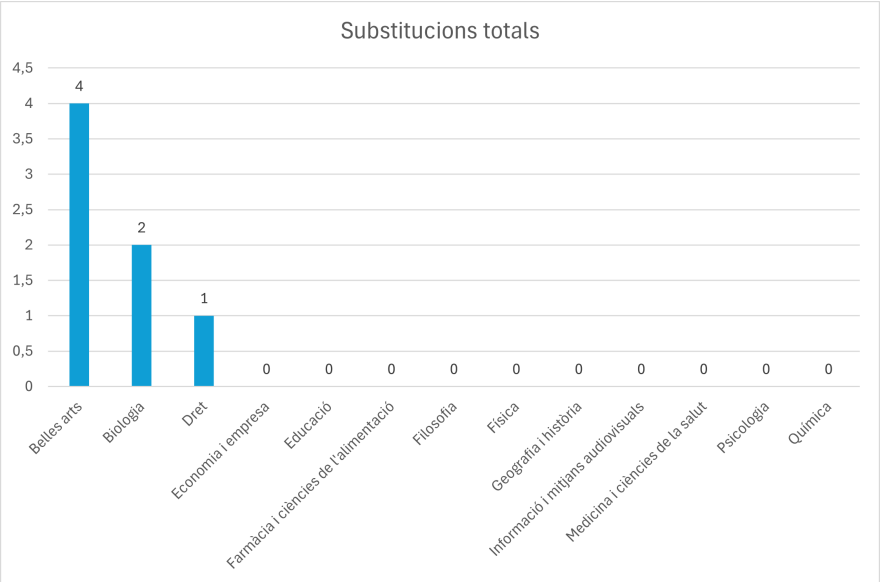


Figura 9: Quantitat de substitucions totals (y) segons la facultat (x).

Apèndix C:

Llista de grups dels graus amb vocabulari similar

Grup 1: Arxivística.txt, Biblioteconomia.txt, Informació-i-documentació.txt

Grup 2: Riscos-naturals.txt, Meteorologia.txt

Grup 3: Edició-web.txt, Informàtica.txt

Grup 4: Hormones.txt, Crononutrició.txt

Grup 5: Zoologia.txt, Fonaments-de-podologia.txt

Grup 6: Astrofísica.txt, Ecologia.txt

Grup 7: Educació-musical.txt, Palinologia.txt

Grup 8: Fitopatologia.txt, Botànica.txt

Grup 9: Malalties-plantes.txt, Ciències-dels-aliments.txt

Grup 10: Odontologia.txt, Podologia.txt

Grup 11: belles-arts.txt, Geologia.txt, Llibre-manuscrit.txt

Grup 12: Semiologia.txt, Anatomia patològica.txt

Grup 13: Nutrició-i-dietètica.txt, Infermeria.txt

Grup 14: Farmàcia-galènica.txt

Grup 15: Auditoria.txt, Comptabilitat.txt, NIIF.txt

Grup 16: Fiscalitat.txt, Economia-i-empresa.txt, Dret.txt

Grup 17: Matemàtiques.txt, Estadística.txt

Grup 18: Educació-física.txt, Didàctica.txt, Didàctica-de-la-llengua.txt

Grup 19: Psicologia-social.txt, Filosofia.txt, ciència-política.txt

Grup 20: Genètica.txt, Bioquímica.txt, Bioquímica-clínica.txt, Biologia-cel·lular-.txt

Grup 21: Física.txt, Astronomia.txt

Grup 22: Materials.txt, Química.txt, Enginyeria-química.txt, Tècniques-instrumentals.txt

Apèndix D:
Codis i corpus

Enllaç al repositori GitHub amb els codis i els corpus emprats per fer aquest treball: https://github.com/clarapuigventos/TFG.git.