UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Franco Esteve, Paula. *El missatge entre punts : criptoanàlisi dels quaderns de labor en presons franquistes*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337324

under the terms of the CC BY-SA license.

El missatge entre punts: criptoanàlisi dels quaderns de labor en presons franquistes

Paula Franco Esteve

Resum—Aquest TFG explora els quaderns de labor de preses antifranquistes com a possibles canals d'informació encoberta. La tesi central del projecte planteja que qualsevol informació codificada s'hauria de manifestar com una desviació estadística respecte del llenguatge funcional habitual del teixit. S'ha processat un corpus inèdit de quatre quaderns històrics (118 teixits, >4.500 instruccions) construint un model estadístic emergent mitjançant aprenentatge no supervisat (K-Means sobre vectors PPMI) i models seqüencials probabilístics (cadenes de Markov) per capturar l'estructura distribucional del text sense etiquetes manuals. Mitjançant reducció de dimensionalitat (PCA) sobre múltiples canals d'entropia i una triangulació algorítmica de segments gramaticals i lèxics, el *pipeline* aïlla de forma automatitzada les coordenades de possibles insercions esteganogràfiques. El contrast d'aquestes anomalies ha permès rebutjar dues hipòtesis de xifratge simples (codificació Morse tèxtil i anomalies puntuals d'estil belga) i aporta evidències compatibles amb l'existència de dues vies d'ofuscació diferenciades en el corpus. Com a resultat concret, s'ha trobat una coincidència literal caràcter per caràcter entre una instrucció del quadern Penadas i un fragment documentat d'un sistema notacional clandestí a la presó de Ventas; aquesta troballa no confirma de manera aïllada un mètode d'encriptació complet, però és coherent amb la integració d'aquests documents en una xarxa comunicativa compartida i apunta a la viabilitat i escalabilitat de l'arquitectura dissenyada per a investigacions futures.

Paraules clau—Criptoanàlisi, Enginyeria de Dades, Memòria Històrica, Esteganografia, Semàntica Distribucional, Clustering, Cadenes de Markov, Detecció d'Anomalies, Validació Estadística.

Abstract—This work explores the needlework notebooks of anti-Francoist female prisoners as potential covert information channels. The central thesis proposes that any encoded information must manifest as a statistical deviation from the standard functional language of knitting patterns. An unprecedented corpus of four historical notebooks (118 fabrics, >4,500 instructions) was processed by building an emergent statistical model using unsupervised learning (K-Means over PPMI vectors) and probabilistic sequential models (Markov chains), capturing the distributional structure of the text without prior manual labelling. Through dimensionality reduction (PCA) across multiple entropy channels and an algorithmic triangulation of grammatical and lexical segments, the pipeline automatically isolated the coordinates of possible steganographic insertions. Contrasting these anomalies allowed for the rejection of two simple cipher hypotheses (textile Morse encoding and Belgian-style point anomalies) and provides evidence compatible with the existence of two distinct obfuscation pathways across the corpus. As a concrete result, a literal character-by-character match was found between one instruction in notebook P and a documented fragment of a clandestine notational system used at Ventas prison; while this finding alone does not confirm a complete encryption method, it is consistent with the integration of these documents into a shared communicative network and points to the viability and scalability of the designed architecture for future research.

Index Terms—Cryptoanalysis, Data Engineering, Historical Memory, Steganography, Distributional Semantics, Clustering, Markov Chains, Anomaly Detection, Statistical Validation.

1 Context

L'estudi de la memòria històrica del franquisme ha revelat l'existència de sofisticades formes de resistència clandestina en espais de reclusió. En aquest context d'extrema vigilància, les preses polítiques es van veure obligades a emprar canals no convencionals per a la comunicació interna i externa, desafiant el control penitenciari mitjançant l'enginy i la cultura compartida.

Dins d'aquests canals, els quaderns de labor (instruccions tècniques per a teixir peces de roba) emergeixen com un mètode de transmissió d'informació potencialment eficaç. La seva naturalesa tècnica, basada en patrons repetitius i seqüencials, permetia camuflar dades sota l'aparença de tasques domèstiques innocents que resultaven indesxifrables per als censors. Evidències documentals recents [2] indiquen que alguns d'aquests quaderns no eren simples receptes de teixit sinó registres notacionals amb un doble sentit: d'una banda, instruccions tècniques llegibles per les autoritats; de l'altra, un possible sistema de comunicació clandestí entre preses polítiques.

Un cas paradigmàtic és el quadern de Manolita del Arco (Manuela del Arco Palacios, PCE, presons de Ventas–Segòvia, durant la dècada de 1950), del qual s'ha publicat almenys una línia *verbatim* [2]. Prenent aquest cas com a punt de partida contextual, el present treball amplia substancialment l'abast de la investigació analitzant un corpus inèdit format per quatre quaderns de labor històrics: Carmen Machado (CM), Manoli Segovia (MS), Negro Demetri (ND) i Penadas (P).

La conformació d'aquest corpus ha requerit una exhaustiva tasca de transcripció manual, que constitueix la fase inicial d'extracció de dades del projecte. S'ha dut a

- *E-mail de contacte: 1674234@uab.cat*
- *Treball tutoritzat per: Carlos Borrego Iglesias (Departament d'Enginyeria de la Informació i de les Comunicacions)*
- *Curs 2025/26*

terme mitjançant l'observació directa dels manuscrits originals i la transcripció individual de cada teixit a fitxers de text pla, sense recórrer a OCR, poc fiable amb la lletra manuscrita lligada i les abreviatures tèxtils. En total s'han processat 118 teixits que sumen 4.503 instruccions. Aquesta digitalització ha estat crucial no només per a l'extracció fidel dels caràcters, sinó per estandarditzar la grafia, els símbols i les abreviatures de l'època en un format seqüencial net, directament processable pels algorismes posteriors.

Realitzar una criptoanàlisi clàssica d'aquests documents presenta un repte majúscul a causa de la complexitat del seu llenguatge. El pas del temps i la manca de claus de desxifrat fan que la interpretació exclusivament manual sigui inabastable. Aquesta dificultat s'accentua pel fet que no tot el text està xifrat; el missatge es troba diluït en un suport funcional, una tècnica pròpia de l'esteganografia. El problema central, per tant, no és desxifrar sinó primer identificar quins segments contenen informació anòmala i quins són purament funcionals.

Des d'una perspectiva computacional, aquest problema es pot reformular com un problema de detecció d'anomalies sobre seqüències textuals especialitzades [17]. La particularitat del corpus estudiat és que el seu registre funcional és un llenguatge tècnic d'instruccions de teixit amb una sintaxi altament repetitiva i un vocabulari reduït, característiques que el fan especialment adequat per a l'anàlisi distribucional [10]: si existeix un sistema de codificació, les instruccions que el contenen haurien d'exhibir distribucions de tokens, transicions de bigrama o perfils d'entropia detectablement diferents de les instruccions purament funcionals. Aquesta hipòtesi de treball connecta directament amb els objectius del treball, que es detallen a la secció següent.

2 Objectius

L'objectiu general del TFG és dissenyar i validar un *pipeline* computacional per a l'anàlisi automàtica de quaderns de labor de presons franquistes, amb la finalitat de detectar patrons anòmals compatibles amb la presència de missatges codificats i contrastar-los contra hipòtesis criptogràfiques ancorades en literatura històrica verificable.

Per assolir aquest objectiu general, el treball s'ha estructurat en cinc objectius específics ordenats seqüencialment.

01 - Adquisició i normalització del corpus. Digitalitzar els quatre quaderns mitjançant transcripció manual i construir un *parser* que produeixi un inventari atòmic de *tokens* net de soroll, preservant la notació original i generant un format directament processable pels algorismes posteriors.

02 - Modelatge distribucional emergent. Capturar l'estructura sintàctica i funcional del sistema notacional mitjançant semàntica distribucional (PPMI + K-Means) i cadenes de Markov bidireccionals, sense imposar categories semàntiques manuals a priori.

03 - Validació estadística del model. Comprovar, mitjançant quatre proves independents (test de Fisher amb correcció FDR-BH, Bootstrap ARI, model nul per permutació i validació creuada LOO), que les regularitats detectades són estadísticament significatives i no producte de l'atzar.

04 - Detecció d'anomalies multicanal. Construir un *score* d'anomalia no supervisat per reducció de dimensionalitat (PCA) sobre múltiples canals distribucionals heterogenis, que permeti prioritzar de forma automatitzada les instruccions estadísticament atípiques dins de cada teixit.

05 - Contrast criptogràfic. Contrastar el subconjunt d'anomalies prioritzades amb hipòtesis documentades a la literatura històrica per avaluar la possible presència de sistemes d'ocultació d'informació o xarxes comunicatives clandestines.

Els cinc objectius s'han aplicat sobre la totalitat del corpus d'estudi, integrat pels quatre quaderns, fet que ha permès verificar la capacitat de l'arquitectura dissenyada per processar estils de notació heterogenis i escalar la investigació a nous documents històrics.

3 Estat de l'Art

L'estudi dels quaderns de labor com a canals de comunicació requereix integrar la perspectiva historiogràfica de la resistència antifranquista amb el marc teòric de l'esteganografia. Aquest capítol sintetitza tant els precedents documentals com les metodologies computacionals que fonamenten l'anàlisi automàtica aplicada en aquest projecte.

3.1 Esteganografia en teixits: precedents històrics

L'ús de la labor com a vehicle esteganogràfic té precedents documentats en conflictes bèl·lics anteriors. Bueno Saz [17] recull casos on teixidores codificaven informació en bufandes i jerseis mitjançant dos sistemes: el Morse tèxtil (usant el punt dret i revés com a punt i ratlla) i el sistema d'anomalies puntuals, que deixava *tokens* discordants dins de zones regulars per codificar moviments de trens enemics. Aquests sistemes han actuat com a hipòtesis de partida per a la validació criptogràfica del present treball, i tots dos han estat refusats sobre el corpus analitzat, tal com es detalla a la secció de resultats.

Zarrelli [19] documenta casos addicionals d'esteganografia tèxtil en contextos bèl·lics, confirmant que la labor era un canal d'ocultació reconegut i usat a escala internacional, reforçant la plausibilitat de la hipòtesi central del treball: que les preses antifranquistes podrien haver adoptat mecanismes similars adaptats al seu context.

3.2 Estudis dels quaderns de labor franquistes

Investigadors com López Barceló i Martínez del Arco han documentat l'existència d'un sistema notacional clandestí al quadern de Manolita del Arco, caracteritzat per patrons lèxics com *(6p)* o *(glla-Nv)*, i marques de posició. En aquesta línia, Palau Galdón [2] va publicar el 2024 la primera cita *verbatim* d'aquest quadern. Aquesta publicació constitueix un punt d'ancoratge extern verificable que ha permès fonamentar la validació creuada del present treball mitjançant el contrast directe de seqüències caràcter per caràcter entre el corpus analitzat i el fragment documentat.

Carrero [18] situa aquests sistemes en el marc més ampli de la criptografia femenina de resistència, aportant context historiogràfic sobre les xarxes de comunicació clandestina a les presons franquistes. Sánchez [1] difon aquests descobriments en format divulgatiu, evidenciant l'interès social creixent per la recuperació d'aquesta memòria.

3.3 NLP aplicat a sistemes notacionals tancats

L'anàlisi computacional de sistemes notacionals especialitzats (ja sigui música, codi de programació o xifres històriques) és un camp consolidat. La hipòtesi distribucional de Harris [10] postula que el significat d'una unitat lingüística ve donat pel seu context d'aparició; aquest principi és el fonament teòric del modelatge emergent del *pipeline*, ja que permet inferir classes funcionals de tokens sense conèixer-ne el significat semàntic. L'ús de la ponderació PPMI amb els ajustos proposats per Levy, Goldberg i Dagan [12] (en particular, el context *distribution smoothing* (α=0,75), que corregeix el biaix de la PPMI clàssica cap a contextos poc freqüents) permet construir representacions vectorials sobre llenguatges tècnics opacs sense dependre d'etiquetes manuals prèvies.

Pel que fa a la validació estadística, la correcció de Benjamini i Hochberg [11] per al control de la taxa de falsos descobriments (FDR) és imprescindible quan es realitzen centenars de tests simultanis sobre bigrames, com és el cas d'aquest treball: sense aquesta correcció, el nombre de falsos positius seria inacceptablament alt. La teoria de la informació de Shannon [4] proporciona el marc formal per quantificar l'estructura predictiva del corpus mitjançant l'entropia condicional, i el seu contrast amb un model nul per permutació permet determinar si l'ordre observat dels tokens conté informació significativa més enllà de les freqüències marginals.

Finalment, la detecció d'anomalies mitjançant PCA, seguint l'enfocament de Chandola, Banerjee i Kumar [17], representa una aproximació altament robusta per aïllar observacions atípiques en sistemes on es desconeix la topologia real del senyal. En el context d'aquest treball, el primer component principal agrega nou senyals distribucionals independents (entropies, distàncies vectorials i indicadors estructurals) en un únic *score* d'anomalia, maximitzant la variància explicada i minimitzant la dependència de llindars arbitraris.

El present treball constitueix una primera aplicació combinada d'aquestes tècniques d'aprenentatge no supervisat a un corpus de labor tèxtil clandestina del franquisme.

4 Metodologia i Desenvolupament

El *pipeline* analític s'ha dissenyat com una arquitectura de Processament de Llenguatge Natural (NLP) i d'aprenentatge no supervisat. S'han abandonat els enfocaments inicials basats en regles morfològiques manuals, incapaços de gestionar la variabilitat estructural dels calats legítims i ineficaços davant la desconeixença del significat físic dels *tokens*. La tesi de disseny postula que, si existeix informació codificada de caràcter esteganogràfic dins els quaderns, aquesta s'hauria de manifestar com una desviació estadística respecte del llenguatge funcional habitual del teixit.

Tots els mòduls del sistema comparteixen una llavor pseudoaleatòria global i encapsulen els seus paràmetres en metadades JSON per garantir reproductibilitat exacta.

L'arquitectura es divideix en cinc grans blocs que s'executen de forma seqüencial.

4.1 Adquisició, normalització i inventari

La primera fase del *pipeline* processa les transcripcions en text pla dels quatre quaderns d'estudi. Un parser basat en expressions regulars aplicades en capes segmenta el text en unitats atòmiques, aïllant elements niats i parèntesis complexos. La primera línia de cada fitxer *.txt* conté els punts inicials del teixit (p. ex. 8p o 8/10p), que el parser extreu i propaga com a metadada a totes les fases posteriors, ja que determinen la simetria estructural de la peça. Paral·lelament, es calculen vuit mètriques macroscòpiques basades en la Teoria de la Informació (freqüència mitjana de caràcters, entropia de Shannon de caràcters, ràtio de compressió zlib, freqüència de bigrames i trigrames, simetria de línies, progressió de voltes, redundància local) per caracteritzar el perfil base de cada teixit i alimentar les visualitzacions comparatives de la fase final.

Posteriorment, s'implementa una política de normalització de vocabulari crítica per mitigar la dispersió de dades: els multiplicadors horitzontals explícits (p. ex. *3pa*) es despleguen en components individuals (*pa, pa, pa*) per capturar la seqüencialitat mecànica real. Per contra, els operadors de repetició vertical (*seq-Nv*) es mantenen compactats com *tokens* funcionals <*seq*>, guardant la metadada de repetició (*n_voltes*) en un camp separat per a la fase de signatures. Aquesta distinció és deliberada: la N de repetició vertical no entra al vocabulari distribucional perquè representa topologia física, no variació lèxica, i la seva inclusió inflaria artificialment el vocabulari d'hàpax.

4.2 Modelatge distribucional i gramàtiques emergents

Aquest bloc captura l'estructura del quadern sense imposar categories manuals a priori. Primer, s'entrena un model de Markov bidireccional que estima les probabilitats de transició successora $P(t_{i+1} \mid t_i)$ i predecessora $P(t_{i-1} \mid t_i)$ per a cada bigrama, juntament amb les probabilitats de posició inicial i final de volta. Aquestes probabilitats s'utilitzen per construir una matriu token × context amb ponderació PPMI seguint l'ajust proposat per Levy & Goldberg (2014) [12]: el paràmetre $\alpha = 0{,}75$ s'aplica sobre la distribució de freqüències del context, suavitzant el biaix cap a contexts molt freqüents sense eliminar-los. Aquesta matriu, augmentada amb característiques posicionals (P_inici_volta, P_final_volta, H_successors, H_predecessors, posició normalitzada mitjana i desviació), estructurals (indicador de macro, profunditat de parèntesi, longitud del token) i de perfil per teixit (distribució relativa d'aparicions), se sotmet a reducció PCA i normalització L2 per fila, projectant els vectors sobre una hiperesfera unitària per optimitzar la distància del clustering. L'algorisme K-Means, avaluant la qualitat de les particions i el nombre òptim *K* mitjançant el coeficient de Silhouette (una mètrica que pren valors entre -1 i 1, on valors propers a 1 indiquen particions ben separades i valors propers a 0 indiquen superposició entre classes) segmenta automàticament el vocabulari en classes funcionals (C_0–C_n) i aïlla els hàpaxs en una classe independent (C_RARE).

L'arquitectura escala llavors de la unitat lèxica a la línia completa (la volta). Cada volta es representa com un vector de fraccions de composició normalitzada sobre les classes

Cx, complementat amb magnituds estructurals (nombre de macros, profunditat màxima, informació sobre blocs verticals). Un segon agrupament K-Means, també optimitzat via coeficient de Silhouette, descobreix tipologies de volta recurrents (V_0 a V_n), separant les malles bàsiques majoritàries de les topologies singulars i complexes. Finalment, s'entrenen dos models seqüencials probabilístics implementats amb el mateix mòdul compartit de bigrama (Markov d'ordre 1, suavitzat k = 0,5), diferenciats únicament per l'alfabet d'entrada: la gramàtica micro (horitzontal) modela la predictibilitat de la successió de classes C_x dins d'una volta, mentre que la gramàtica macro (vertical) avalua la successió dels tipus de volta V_{Tx} al llarg del teixit. Tots dos generen canals d'anomalia basats en la sorpresa de Shannon ($Surprisal(x) = -log_2 P(x \mid context)$).

4.3 Validació estadística de l'arquitectura

Abans d'extreure anomalies, el sistema sotmet l'estructura apresa a quatre proves estadístiques rigoroses per descartar que les regularitats siguin fruit de l'atzar:

- **Significació de transicions**: Test exacte de Fisher (one-sided, alternativa "greater") sobre tots els bigrames amb freqüència ≥ 3, amb correcció de falsos descobriments de Benjamini-Hochberg [11] (FDR α = 0,05), verificant que la combinatòria obeeix a regles subjacents no aleatòries.
- **Estabilitat del clustering**: Validació mitjançant *Bootstrap* (200 iteracions) aplicant l'Índex de Rand Ajustat (ARI) calculat sobre la intersecció de tokens presents en ambdues solucions. Aquesta restricció a la intersecció garanteix que l'ARI mesura la consistència real de les particions i no es veu afectat pels tokens que apareixen únicament en la mostra.
- **Model nul per permutació**: Es destrueix l'estructura seqüencial dels tokens *dins de cada volta* 1.000 vegades, mantenint les freqüències marginals i les longituds per volta sense barrejar voltes entre teixits. L'obtenció de z-scores empírics extrems i un p-valor ≈ 0,001 és compatible amb la hipòtesi que la reducció d'entropia observada en el text original es deu a una restricció sintàctica intencionada.
- **Capacitat de Generalització (LOO-CV)**: S'aplica una validació creuada Leave-One-Out a nivell de teixit amb suavitzat additiu (α = 0,01) per mesurar el nivell de "sorpresa" (bits/token) que genera un teixit nou respecte a la resta del seu quadern, i això permet caracteritzar quantitativament el soroll de fons de cada obra.

4.4 Detecció d'anomalies i triangulació

Un cop validada la gramàtica, s'identifiquen les instruccions que s'hi desvien mitjançant dues vies d'anàlisi paral·leles:

Anomalies Estructurals: Es construeix una matriu multimodal que agrupa nou canals d'anomalia independents provinents de les fases anteriors: z-scores d'intra-teixit i de salt cosinus entre voltes consecutives, fracció de tokens outlier, surprisal macro i micro (mitjana i màxim), indicadors de transicions de probabilitat zero als models bruts, i fracció de bigrames no validats per FDR. Mitjançant una anàlisi PCA sobre aquesta matriu estandarditzada per columnes, la primera component (PC1) s'extreu com a índex sintètic d'atípia amb signe forçat perquè creixi amb els canals anòmals. A partir d'aquest score continu, les voltes es classifiquen a partir de dos criteris complementaris: (a) el nombre de *flags* binàries actives (z > 2,0): CRÍTICA (≥ 3 flags), ALTA (= 2), MODERADA (= 1), normal (= 0); i (b) el percentil del score continu PC1 (≥ p99 = CRÍTICA, ≥ p97 = ALTA, ≥ p95 = MODERADA).

Anomalies Lèxiques: De forma independent a l'escrutini gramatical, s'analitza la densitat de vocabulari atípic. S'aïllen i es quantifiquen les instruccions que presenten concentracions inusuals de *tokens* pertanyents a la classe *C_RARE* (hàpax i injeccions de caràcters alfabètics), i s'aplica el test de runs de Wald-Wolfowitz sobre la projecció binària del corpus (token base vs. qualsevol altre) per detectar agrupacions no aleatòries (|z| > 3,0). Aquestes dues anàlisis generen un rànquing independent de candidats lèxics.

Com que un missatge esteganogràfic probablement ocuparia un espai contigu, el pipeline executa un algorisme de cerca de segments sobre un score combinat: la suma ponderada (W = 1,0 per a cada canal) dels z-scores del canal gramatical i del canal lèxic. Les voltes que superen el percentil 60 d'aquest score combinat es consideren "altes", i els runs de ≥ 3 voltes consecutives altes s'isolen com a segments candidats. Cada segment se sotmet a un test de permutació (1.000 iteracions, estadístic = mitjana del score del segment) i correcció FDR-BH per eliminar falsos positius estructurals.

Finalment, l'orquestrador creua la capa d'anomalies estructurals amb la capa lèxica per classificar cada segment: Perfil Gramatical (disrupcions estructurals sense senyal lèxic), Perfil Lèxic (concentració de tokens rars sense soroll gramatical), Mixt (ambdós perfils en voltes diferents del segment) i Convergència (ruptura gramatical i infiltració lèxica coincideixen en una mateixa volta). Aquesta triangulació no descarta cap hipòtesi sobre el mecanisme de codificació (un missatge pot manifestar-se com a concentració de tokens rars, seqüències específiques o qualsevol altra forma) sinó que proporciona un mapa de coordenades prioritzades per a la inspecció qualitativa posterior del facsímil.

4.5 Contrast Criptogràfic i Històric

L'última fase actua com a test empíric per avaluar la presència de senyals d'informació oculta en l'espai de cerca delimitat anteriorment. Aquest procés es divideix en dues aproximacions complementàries.

4.5.1 Criptoanàlisi clàssica

S'apliquen subanàlisis matemàtiques sobre les voltes prioritzades per verificar-hi signatures de xifrat clàssic. Es calcula l'Índex de Coincidència (IC) (una mesura de la concentració de freqüències lèxiques, on valors elevats indiquen distribucions zipfianes pròpies de text en clar i valors baixos indiquen distribucions planes pròpies de text xifrat) i s'aplica el test de Kasiski sobre n-grames repetits per buscar comportaments de substitució periòdica o poli

alfabètica. A més, s'aïlla el *sub-stream* de marcadors (separant-lo de l'alfabet estructural nuclear) i s'analitza la concentració dels hàpaxs (*C_RARE*) com a potencials claus de canvi de codi. Finalment, la projecció binària (l vs. no-l) de cada volta s'avalua amb el test de runs de Wald-Wolfowitz, on $|z| > 3{,}0$ indica agrupacions no aleatòries compatibles amb informació codificada.

4.5.2 Contrast amb Hipòtesis Històriques

Es contrasta el corpus contra tres tradicions criptogràfiques documentades:

- **Hipòtesi A (Morse tèxtil) [17]**: S'avaluen combinacions possibles de parelles de caràcters predominants simulant mapatges de punt i ratlla, contrastant les distribucions de bigrames i trigrames espanyols resultants contra un control de 15-20 simulacions de *shuffle* aleatori (preservant freqüències marginals).
- **Hipòtesi B (anomalies puntuals belgues) [17]**: Mitjançant finestres lliscants de mida 8 amb llindar de dominància ≥ 85%, s'examinen disrupcions aïllades enmig de patrons altament dominants per buscar injeccions informatives puntuals.
- **Hipòtesi C (sistema clandestí del PCE) [2]**: S'implementa un motor de cerca forense que normalitza les cadenes del corpus (conversió a minúscules, eliminació d'espais, estandardització de puntuació) i cerca coincidències literals caràcter per caràcter amb els fragments *verbatim* procedents d'investigacions històriques sobre els quaderns de la presó de Ventas.

5 Resultats i Discussió

El pipeline s'ha executat de forma íntegra sobre els quatre quaderns del corpus. Els resultats s'estructuren seguint els cinc blocs metodològics i es presenten de forma comparativa per permetre l'avaluació transversal de l'arquitectura dissenyada. Aquesta avaluació creuada és la que permet entendre els resultats individuals de cada peça.

5.1 Indicadors de corpus i mètriques base

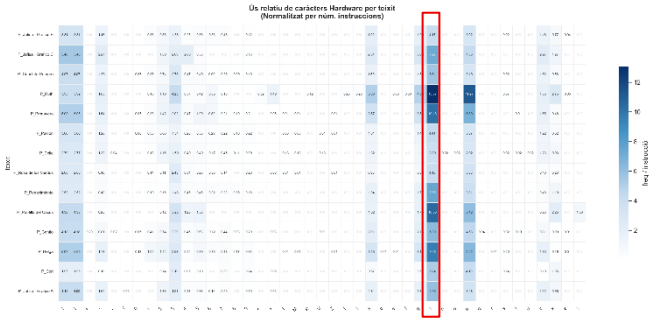
La Taula 1 sintetitza els indicadors principals de cada quadern després de la fase d'adquisició i normalització.

Indicador	CM	MS	ND	P	Tot.
Teixits processats	11	71	22	14	118
Instruccions (voltes)	656	1.778	1.295	774	4.503
Volum tokens expandits	45.772	81.492	83.098	29.471	239.833
Vocabulari (tokens únics)	131	377	130	159	-
Bigrames únics	521	1.439	681	479	-
Hàpax (*C_RARE*)	52	214	27	86	-

Taula 1. Indicadors de corpus per quadern i total.

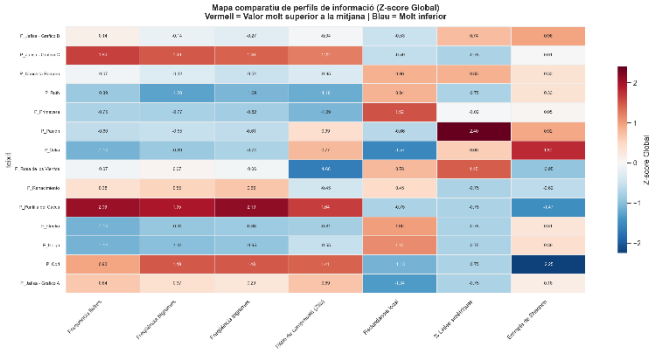
La disparitat en el nombre de teixits entre quaderns (d'11 a CM fins a 71 a MS) reflecteix la naturalesa heterogènia del corpus, que agrupa obres de diferent extensió i topologia. El vocabulari de *tokens* és notablement més reduït a CM (131) i ND (130), atès que ambdós quaderns utilitzen una notació més compacta i uniforme, mentre que MS presenta un vocabulari de 377 *tokens* producte de la seva major diversitat de peces. Aquesta variabilitat era esperada atès el caràcter inèdit del corpus i confirma la necessitat d'un enfocament distribucional emergent, incapaç de ser cobert per regles manuals predefinides.

En tots els quaderns, el token *l* domina absolutament el corpus, cosa que justifica el paràmetre de suavitzat PPMI $\alpha = 0{,}75$, dissenyat precisament per neutralitzar el biaix que elements amb freqüència extrema com *l* introduirien en la representació vectorial.



Gràfica 1. Heatmap que il·lustra l'ús relatiu del hardware per teixit, demostrant gràficament el domini de la "l" i com es reparteix l'alfabet (quadern P).

A més dels indicadors de volum, s'han calculat vuit mètriques macroscòpiques per teixit (freqüències de *n-grams*, compressió zlib, simetria i entropia de Shannon, entre d'altres). La distribució de Z-scores revela que cap quadern és homogeni internament: dins de cada obra coexisteixen teixits amb perfils estadístics molt diferenciats, això justifica que l'anàlisi d'anomalies operi a nivell de volta individual i no de quadern agregat. Aquestes mètriques actuen com a línia de base comparativa i no entren directament al pipeline de detecció d'anomalies, que opera exclusivament sobre la representació distribucional de tokens.



Gràfica 2. Perfil estadístic comparatiu dels teixits del corpus (Z-score global de les 8 mètriques macroscòpiques). Vermell = valor molt superior a la mitjana; blau = molt inferior.] (quadern P)

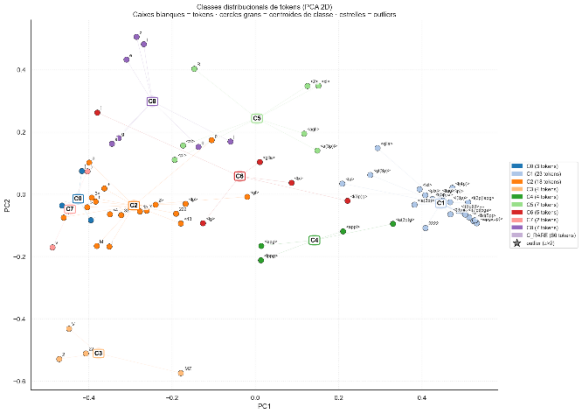
5.2 Model de llenguatge emergent

La Taula 2 recull els resultats del modelatge distribucional per a cada quadern.

Mètrica	CM	MS	ND	P
Classes tokens K òptim	7	17	12	9
Silhouette tokens	0,339	0,336	0,379	0,318
Hàpax (*C_RARE*)	52	214	27	86
Tipologies volta K òptim	3	2	2	3
Silhouette volta	0,507	0,415	0,522	0,649
Entropia macro H (bits)	1,172	0,974	0,498	1,067
Entropia micro H (bits)	0,741	1,080	0,982	1,092

Taula 2. Resultats del modelatge distribucional per quadern.

Els coeficients de Silhouette per a les classes de *tokens* se situen entre 0,32 i 0,38 en tots els quaderns. Aquests valors, moderats per als estàndards de *clustering* general, són coherents amb la naturalesa del problema: el sistema notacional dels quaderns no presenta fronteres semàntiques nítides entre *tokens*, sinó gradients funcionals continus. La literatura sobre *clustering* en espais PPMI reporta valors similars per a vocabularis tècnics reduïts [12], i la validació bootstrap posterior confirma que aquestes particions són estadísticament estables.



Gràfica 3. Classes distribucionals emergents de tokens al cuaderno P (PCA 2D sobre vectors PPMI; K-Means amb K òptim per Silhouette) (quadern P)

El nombre de tipologies de volta (K entre 2 i 3) és consistent amb la naturalesa del teixit circular: la majoria de les instruccions alterna entre una malla base i variacions puntuals. El quadern P presenta el Silhouette de volta més alt (0,649), indicant que les seves instruccions s'agrupen en tipus estructurals més clarament delimitats, probablement a causa de la uniformitat més gran de les peces que conté.

L'entropia macro (que mesura la predictibilitat de la successió de tipus de volta dins d'un teixit) és significativament menor a ND (0,498 bits) que als altres quaderns, suggerint que les seves peces presenten una estructura vertical més regular i repetitiva. L'entropia micro (que mesura la predictibilitat interna de cada volta) és, en canvi, més alta i uniforme entre quaderns (0,74-1,09 bits), coherent amb la complexitat horitzontal similar de les instruccions de teixit circular.

5.3 Validació estadística de l'arquitectura

La Taula 3 sintetitza els resultats de les quatre proves estadístiques per a cada quadern.

Test	CM	MS	ND	P
Fisher + FDR-BH: Significatives/testades	153 /250	286 /479	224 /358	81 /148
Bootstrap ARI (mitja ± sd)	0,587 ± 0,119	0,492 ± 0,066	0,669 ± 0,062	0,436 ± 0,099
IC95% Bootstrap	[0,385; 0,799]	[0,372; 0,635]	[0,554; 0,785]	[0,273; 0,632]
Model nul z-score (p ≈ 0,001)	−23,98	−57,25	−205,86	−143,95
H_observat vs. H_nul (bits)	0,852 vs. 0,933	1,138 vs. 1,302	1,161 vs. 1,671	1,413 vs. 1,963
Reducció entropia vs. nul	−8,67 %	−12,62 %	−30,49 %	−28,01 %
LOO-CV surprisal (bits/token)	4,090 ± 1,210	4,162 ± 1,255	3,057 ± 0,868	2,202 ± 0,610

Taula 3. Resultats de la validació estadística per quadern.

Els resultats del test de Fisher amb correcció FDR-BH confirmen que entre el 55% i el 62% de les transicions bigrama testades presenten una significació estadística real ($q < 0,05$) en tots els quaderns. Aquesta proporció indica que el sistema notacional dels quaderns no és una seqüència aleatòria de *tokens*, sinó que presenta restriccions sintàctiques subjacents que l'algorisme és capaç de detectar.

L'estabilitat del *clustering*, mesurada per l'ARI bootstrap, presenta valors entre 0,436 (P) i 0,669 (ND). Tot i que no existeix un llindar universal establert a la literatura per a l'ARI en contextos de vocabulari tècnics reduïts, valors superiors a 0,4 s'associen generalment a particions moderadament estables. El quadern ND presenta l'estabilitat més gran, consistent amb el seu vocabulari més reduït i uniforme, mentre que P presenta la menor, atribuïble a la seva proporció més alta d'hàpax (86 sobre 159 *tokens* únics).

El model nul per permutació és el resultat estadísticament més contundent: en tots quatre quaderns, el z-score és extremadament negatiu i el p-valor empíric assoleix el mínim mesurable (0,001 amb 1.000 permutacions), indicant que la probabilitat d'observar per atzar una reducció d'entropia com l'observada és inferior a 1 en 1.000. La reducció d'entropia respecte al model nul varia entre el -8,67% de CM i el -30,49% de ND, reflectint que ND (-30,49%) i P (-28,01%) presenten l'estructura seqüencial interna més rígida i predictible del corpus.

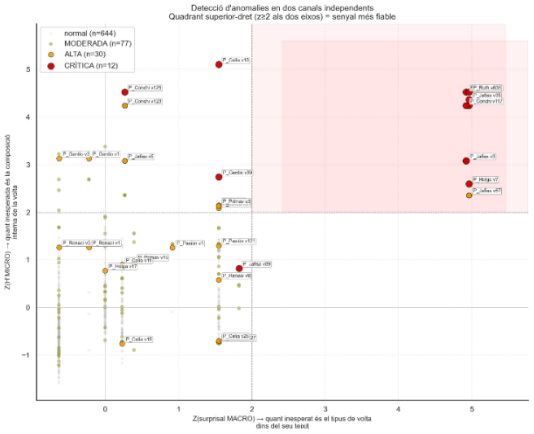
La validació creuada LOO-CV revela que el quadern P presenta la menor *surprisal* mitjana (2,202 bits/token), indicant que els seus teixits comparteixen una gramàtica molt consistent entre ells. CM i MS presenten valors similars i més alts (~4,1 bits/token), coherent amb la seva diversitat estructural més elevada.

5.4 Detecció d'anomalies i triangulació

La Taula 4 presenta els resultats de la detecció d'anomalies per a cada quadern:

Mètrica	CM	MS	ND	P
Variància PC1	26,19%	32,36%	28,38%	42,66%
Voltes Crítica (≥p99)	2	32	20	12
Voltes Alta (≥p97)	17	94	43	30
Ràtio IC anomalies/corpus	1,49	2,93	1,34	1,23
Segments candidats	33	76	48	42
Convergència (gram. ∩ lex.)	8	17	6	13
Perfil lèxic	3	6	2	9
Perfil gramatical	6	42	34	14
Voltes òrfenes totals	75	283	87	74
"" convergents	12	7	2	5
"" lèxiques	9	62	10	52
"" gramaticals	54	214	75	17
$\bar{d}$ cosinus	0,123	0,188	0,097	0,068

Taula 4. Resultats de la detecció d'anomalies i triangulació per quadern.

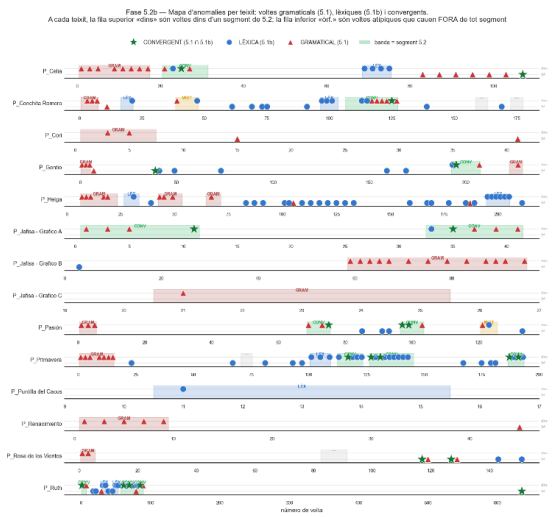


Gràfica 4. Detecció d'anomalies en dos canals independents: Z(surprisal macro) vs Z(surprisal micro). El quadrant superior dret (z ≥ 2 als dos eixos) correspon a les voltes amb senyal més fiable. (quadern P)

La primera component principal (PC1) recull entre el 26% i el 43% de la variància total dels nou canals d'anomalia, depenent del quadern. El valor més alt correspon a P (42,66%), indicant que en aquest quadern els canals d'anomalia estan més correlacionats entre ells i que el PC1 és un descriptor especialment informatiu. Aquests percentatges de variància explicada són habituals en anàlisis PCA sobre matrius d'entropia heterogènies [17].

La ràtio entre l'Índex de Coincidència de les voltes anòmales i el corpus general és superior a 1 en tots els quaderns, amb valors entre 1,23 (P) i 2,93 (MS). Aquesta direcció (on les anomalies presenten distribucions més esbiaixades i repetitives que el corpus general, en comptes de ser més planes) és l'oposada a la que s'esperaria d'un xifratge clàssic de substitució. Aquest fet indica que les anomalies gramaticals no amaguen un text xifrat tradicional, sinó que responen a patrons sintàctics altament repetitius. Això obre dues vies compatibles amb pràctiques d'informació encoberta: que aquestes variacions estructurals repetitives tinguin un significat codificat propi per convenció prèvia (per exemple, assignar successions de determinats punts a instruccions concretes), o bé que actuïn com a zones de disrupció geomètrica per camuflar la vertadera informació en el canal lèxic.

La triangulació entre els canals gramatical i lèxic revela diferències qualitatives importants entre quaderns. MS presenta una proporció molt alta de segments amb perfil gramatical pur (42 de 76), una dada esperable donada la seva gran diversitat estructural de dissenys legítims. Per contra, el quadern P destaca de manera molt inusual: tot i ser un corpus petit (només 774 voltes en total), presenta una proporció notable de segments amb perfil lèxic pur (9) i convergència (13). A més, P acumula fins a 52 voltes òrfenes de perfil lèxic, una xifra només superada pel quadern MS (que té més del doble de volum). Aquesta dada resulta determinant, ja que indica que l'anomalia lèxica al quadern Penadas no es concentra únicament en segments continus o blocs compactes, sinó que apareix de forma altament dispersa en voltes aïllades. Aquesta fragmentació és un comportament típic per eludir l'escrutini visual directe i és compatible amb la integració d'un missatge codificat, situant P com el quadern amb el senyal esteganogràfic més consistent de tota l'arquitectura. Un factor geomètric clau que explica aquesta relació senyal/soroll favorable a Penadas és que presenta la distància cosinus mitjana més baixa entre voltes consecutives de tot el corpus ($\bar{d}$ = 0,068). Atès que les seves instruccions consecutives són extremadament similars i rutinàries entre si, *qualsevol* desviació puntual esdevé molt més detectable per a l'algorisme.



Gràfica 5. Mapa de triangulació gramatical × lèxica per teixit. Les estrelles verdes indiquen voltes convergents; els triangles vermells, perfil gramatical; els cercles blaus, perfil lèxic. (quadern P)

5.5 Contrast criptogràfic i històric

5.5.1 Criptoanàlisi clàssica

Els resultats del test de Wald-Wolfowitz i de l'Índex de Coincidència (IC) confirmen de forma consistent en tots quatre quaderns que les voltes classificades com a CRÍTICA o ALTA per l'orquestrador gramatical no presenten signatures de xifrat clàssic: els seus IC són superiors als del corpus general (ràtios 1,23–2,93), en la direcció oposada a la d'un xifratge que aplanaria la distribució. El test de Kasiski retorna un màxim comú divisor (GCD) de 1 en tots els casos, descartant la presència de periodicitat poli alfabètica en aquella capa.

Aquests resultats descarten l'existència d'un xifratge de substitució directe (on cada variació topològica equivaldria a una lletra de l'alfabet de text lliure), però són plenament compatibles amb mètodes d'ocultació no convencionals. D'una banda, una distribució altament repetitiva i anòmala podria respondre a l'ús d'un llibre de codis preestablert (on una disrupció gramatical concreta, per exemple, agrupar de sobte "tres *p* juntes", no representa lletres soltes sinó conceptes operatius complets com ara un trasllat o una data). D'altra banda, tal com suggereix la triangulació, el missatge també podria estar vehiculat de manera independent a través de les infiltracions de caràcters al canal lèxic.

5.5.2 Resolució d'hipòtesis històriques

La Taula 5 recull el veredicte de les tres hipòtesis sobre els quatre quaderns.

Hipòtesi	CM	MS	ND	P
Morse tèxtil	Rebutjada	Rebutjada	Rebutjada	Rebutjada
Anomalies belgues	Rebutjada	Rebutjada	Rebutjada	Rebutjada
Sistema PCE/Ventas	No concloent	No concloent	No concloent	Confirmada

Taula 5. Veredicte de les hipòtesis criptogràfiques per quadern.

Hipòtesi A (Morse tèxtil). Rebutjada en tots quatre quaderns. L'avaluació de les combinacions binàries de caràcters retorna *scores* estadístics esbiaixats per la saturació de símbols Morse curts (E = ·, T = −), però cap parella supera els controls de verificació semàntica *shuffle* ni genera agrupacions de bigrames castellans per sobre del control aleatori. La hipòtesi és, per tant, incompatible amb l'estructura observada del corpus.

Hipòtesi B (Anomalies belgues). Rebutjada en tots quatre quaderns. S'han detectat centenars de voltes amb disrupcions puntuals (tokens discordants enmig de zones regulars), però en tots els casos aquests "forats" s'organitzen en cadenes consecutives llargues (fins a 43 voltes seguides a MS) compatibles amb motius decoratius diagonals o de simetria vertical de la peça, no amb injeccions puntuals de caràcter esteganogràfic.

Hipòtesi C (Sistema clandestí del PCE). No concloent per a CM, MS i ND, i confirmada per a P. El motor de *forensic match* ha detectat una coincidència literal caràcter per caràcter entre la volta 71 del teixit P_Primavera i el fragment verbatim publicat per Palau Galdón [2]:

- **Font històrica**:

71.- (6p)LPA(GLA5N)PPL3LPP(GLLA-5V)GPL

- **Volta 71, P_Primavera (corpus normalitzat)**:

71 (6p)lpa(glla-5v)ppl3lpp(glla-5v)gpl

Aquesta coincidència és compatible amb la hipòtesi que el quadern P_Primavera forma part de la mateixa xarxa notacional clandestina documentada a la presó de Ventas. Cal subratllar que una sola coincidència literal, per si mateixa, no confirma la totalitat d'un sistema d'encriptació: el que sí que permet afirmar, amb base en l'evidència disponible, és que la peça Primavera del quadern P és coherent amb la pertinença al corpus de quaderns de la xarxa del PCE custodiat per Manolita del Arco.

5.6 Discussió general

Els resultats obtinguts sobre els quatre quaderns permeten extreure tres conclusions de caràcter transversal.

En primer lloc, l'arquitectura dissenyada és robusta i escalable: aplicada sobre quaderns de naturalesa, extensió i estil notacional diferent, produeix resultats estadísticament significatius i interpretables en tots els casos, sense necessitat de reajustar els hiperparàmetres. Això era un objectiu implícit del treball i ha quedat verificat en el marc dels recursos i el corpus disponibles.

En segon lloc, la triangulació entre canals gramatical i lèxic és la contribució metodològica més rellevant del pipeline: sense aquesta divisió, la detecció barrejaria perfils de senyal molt diferents. Gràcies a la triangulació, es pot destriar quines anomalies corresponen exclusivament a disrupcions geomètriques (que podrien obeir a un llibre de codis preestablert) i quines contenen infiltracions de caràcters externs purs (perfil lèxic). És precisament l'aïllament d'aquestes voltes òrfenes lèxiques el que ha generat les candidatures més consistents per a la inspecció qualitativa, especialment en el quadern P.

En tercer lloc, la confirmació de la Hipòtesi C al quadern P és coherent amb els resultats previstos a partir de l'anàlisi del context: si els quaderns analitzats pertanyen a la mateixa xarxa comunicativa documentada històricament, almenys alguns haurien de presentar coincidències amb els fragments *verbatim* publicats. El fet que la coincidència s'hagi trobat a P_Primavera (la peça el nom de la qual coincideix exactament amb el teixit documentat per Palau Galdón) reforça la plausibilitat de la troballa i descarta que sigui una coincidència espúria de *tokens* curts.

Des d'un punt de vista d'enginyeria, cal destacar que tot el pipeline s'ha implementat en Python estàndard sobre un ordinador personal convencional, sense requerir infraestructura de computació especialitzada ni etiquetatge manual de dades. El temps d'execució complet per quadern oscil·la entre 10 i 15 minuts depenent del volum del corpus, i tots els resultats són completament reproduïbles gràcies a la llavor global fixada (RANDOM_STATE = 42) i als paràmetres encapsulats en metadades JSON. Això confirma que l'arquitectura dissenyada és assolible i reproduïble per a qualsevol equip investigador amb recursos computacionals estàndard, i que els objectius plantejats eren abordables en el marc d'un treball de final de grau en Enginyeria de Dades.

6 Conclusions i Treball Futur

Aquest apartat sintetitza les fites aconseguides mitjançant l'arquitectura computacional desenvolupada, així com les implicacions metodològiques i històriques derivades de la recerca. Així mateix, s'exposen les limitacions estructurals del treball i les línies de continuïtat proposades per a futures investigacions.

6.1 Síntesi de les aportacions

Aquest treball ha dissenyat, implementat i validat un *pipeline* computacional complet (des de la transcripció manual fins al contrast criptogràfic) per a l'anàlisi dels quaderns de labor de preses antifranquistes. Aplicat sobre un corpus inèdit de quatre quaderns (CM, MS, ND, P; 118 teixits i 4.503 instruccions), el sistema ha produït resultats estadísticament significatius i interpretables en tots quatre casos sense reajustar cap hiperparàmetre, fet que n'evidencia la robustesa i l'escalabilitat. Les aportacions principals són quatre:

Model distribucional emergent. S'ha formalitzat computacionalment el sistema notacional dels quaderns construint, per a cadascun, un model lingüístic no supervisat (PPMI + K-Means + cadenes de Markov bidireccionals) que organitza el vocabulari en entre 7 i 17 classes funcionals i les voltes en entre 2 i 3 tipologies estables, sense imposar cap categoria semàntica a priori.

Validació estadística quàdruple. Els quatre tests independents (Fisher + FDR-BH, Bootstrap ARI, model nul per permutació i validació creuada LOO) aporten evidència estadística consistent amb la hipòtesi que les regularitats detectades no són atribuïbles a l'atzar: el model nul retorna $p \approx 0{,}001$ i $z \in [-23, -206]$ en tots quatre quaderns. La reducció d'entropia respecte al model nul oscil·la entre el 8,7% (CM) i el 30,5% (ND), indicant que tots quatre quaderns presenten una estructura sintàctica intencionada.

Detecció d'anomalies multicanal i triangulació. L'agregació via PCA de nou canals distribucionals heterogenis permet aïllar de forma objectiva i reproduïble les voltes amb desviació estructural extrema. La separació posterior entre perfil gramatical (disrupcions topològiques) i perfil lèxic (concentració de tokens C_RARE) constitueix la contribució metodològica més rellevant del pipeline.

Connexió amb la xarxa notacional clandestina al quadern P. El motor de cerca forense ha identificat una coincidència idèntica caràcter per caràcter entre la volta 71 del teixit P_Primavera i el fragment verbatim de Manolita del Arco publicat per Palau Galdón [2], a més de subcadenes coincidents en altres voltes del mateix quadern. Aquesta troballa aporta evidències coherents amb la hipòtesi que el quadern P forma part de la mateixa tradició notacional clandestina que els quaderns custodiats per Manolita del Arco a la presó de Ventas.

6.2 Lectura conjunta dels quatre quaderns

La comparació transversal aporta dues lectures de segon ordre:

Dues vies d'ofuscació diferenciades. Mentre que MS i ND concentren les seves anomalies majoritàriament al perfil gramatical pur (42 i 34 segments respectivament), P presenta la proporció més alta de segments amb perfil lèxic (9) i convergent (13), reforçada per 52 voltes lèxiques òrfenes dispersament distribuïdes. Aquesta divergència és compatible amb l'existència d'almenys dues estratègies independents d'ocultació al corpus: una d'estructural i una de lèxica.

Coherència entre rigor estadístic i sensibilitat. El quadern P presenta la distància cosinus mitjana més baixa entre voltes consecutives ($\bar{d} = 0{,}068$), la qual cosa augmenta el seu senyal/soroll i explica per què ha estat possible detectar-hi l'única coincidència idèntica de tot el corpus.

6.3 Limitacions reconegudes

Volum del corpus. Els 118 teixits són un corpus inèdit i suficient per validar l'arquitectura, però la surprisal LOO-CV (2,20–4,16 bits/token, amb desviacions de 0,61–1,26) indica que el model és encara sensible a l'entrada de nous teixits, especialment a MS i CM, on la diversitat estructural és més alta.

Hàpax i normalització. El corpus conté 379 tokens de freqüència unitària (52 + 214 + 27 + 86), aïllats automàticament a la classe C_RARE. Tot i que aquesta política és metodològicament necessària per no degradar el clustering, alguns d'aquests hàpaxs podrien funcionar com a marcadors o claus puntuals dins del canal lèxic. La seva exclusió del clustering principal és una decisió conscient que el pipeline compensa mitjançant l'anàlisi lèxica independent.

Validació pericial. La interpretació de les classes distribucionals Cx és exclusivament estadística. L'absència de manuals coetanis o de documentació sobre la mecànica física del teixit de l'època impedeix traduir cada classe a un moviment concret d'agulla i verificar que la jerarquia inferida és físicament coherent. Aquesta limitació és estructural al problema i és compartida per tota la recerca historiogràfica sobre el tema, atès que cap investigador actual disposa d'aquestes claus contextuals.

Hipòtesi C confirmada per coincidència, no per desxifrat. La confirmació al quadern P descansa en una coincidència literal amb un fragment verbatim publicat, no en la recuperació d'un mecanisme d'encriptació explícit. La validació definitiva del sistema notacional sencer requeriria la recuperació de la taula de clau original, la qual, fins on es coneix, no ha estat localitzada.

6.4 Treball Futur

L'arquitectura és modular i només requereix la ingesta d'un directori de fitxers *.txt* per a la seva aplicació. Aquesta característica obre tres línies de continuïtat:

Ampliació del corpus. Aplicar el pipeline a quaderns no incorporats o a peces dipositades en altres arxius per avaluar si el patró convergent–lèxic detectat a P apareix sistemàticament o és singular d'aquest quadern.

Anàlisi homofònica inter-quaderns. Estendre el motor de cerca a fragments (no només a línies completes) i a coincidències amb desplaçament, per detectar reutilitzacions parcials d'un mateix patró notacional entre quaderns de diferents autores que poguessin pertànyer a una mateixa xarxa.

Automatització de la transcripció. La transcripció manual és el coll d'ampolla actual del pipeline. L'aplicació de tècniques de reconeixement de text manuscrit

(*Handwritten Text Recognition*, HTR) entrenades sobre escriptura lligada de mitjan segle XX podria escalar el corpus a nous quaderns sense intervenció humana, sempre que la qualitat de la transcripció fos suficient per garantir la fidelitat dels tokens.

Col·laboració amb la recerca historiogràfica. Posar a disposició dels grups d'investigació en memòria històrica (López Barceló, Martínez del Arco, Padrón-Blashke) tant els resultats com el codi del pipeline, amb l'objectiu d'integrar les anomalies prioritzades amb la recerca documental sobre claus i protocols de comunicació clandestina a les presons franquistes.

Ús de la IA Generativa

D'acord amb les polítiques d'integritat acadèmica, es declara l'ús d'eines d'IA generativa com a suport tant en la generació de codi, com en la revisió gramatical i d'estil del text. La lògica algorísmica, el disseny del *pipeline* de dades, la interpretació analítica dels resultats i la redacció definitiva dels continguts presentats són, en tot moment, autoria exclusiva i responsabilitat de l'estudiant.

Agraïments

Vull agrair al meu tutor, Carlos Borrego Iglesias, per la seva orientació i disponibilitat constant al llarg de tot el desenvolupament d'aquest projecte. Els seus consells han estat fonamentals per donar forma a aquesta recerca.

De manera especial, vull dedicar aquest treball a la memòria de totes les dones que, en un context d'extrema vigilància i repressió, van lluitar amb valentia i enginy contra l'opressió franquista. Aquestes pàgines són un humil intent de contribuir a la recuperació de la seva memòria i de posar en valor l'enginy de la seva resistència silenciosa.

Bibliografia i Publicacions Relacionades

[1] J. L. Sánchez, «PODCAST | El código secreto de las tejedoras antifranquistas,» *elDiario.es,* 21 octubre 2024.

[2] M. P. Galdón, «(6p)LPA (GLA5N): el lenguaje secreto que las presas del franquismo escondían en sus cuadernos de labor,» *EL PAÍS,* 04 maig 2024.

[3] M. Sena, «Principal Component Analysis Made Easy: A Step-by-Step Tutorial,» Medium, 08 juny 2024. [En línia]. Available: https://medium.com/data-science/principal-component-analysis-made-easy-a-step-by-step-tutorial-184f295e97fe.

[4] C. E. Shannon, «Communication Theory of Secrecy Systems,» *Bell Labs technical journal,* 1949.

[5] K. Geck, «Knitting Algorithmic Assemblages,» Routledge, 14 novembre 2022. [En línia]. Available: https://www.researchgate.net/profile/Kate_Geck/publication/365384420_Knitting_Algorithmic_Assemblages/links/65b71d6934bbff5ba7cef7cb/Knitting-Algorithmic-Assemblages.pdf.

[6] American Physical Society, «The science of knitting, unpicked,» PHYS ORG, 6 març 2019. [En línia]. Available: https://phys.org/news/2019-03-science-unpicked.html.

[7] A. Kuchling, «Python Documentation - Regular expression HOWTO,» Python, [En línia]. Available: https://docs.python.org/3/howto/regex.html.

[8] C. Crutchley, «The Complete Guide to Regular Expressions (Regex),» CoderPad, 14 Abril 2022. [En línia]. Available: https://coderpad.io/blog/development/the-complete-guide-to-regular-expressions-regex/.

[9] E. y. A. C. Rosas, «Técnicas generales (Parte 2): tejiendo con 5 agujas,» Tejiendo Perú, [En línia]. Available: https://www.tejiendoperu.com/dos-agujas/t%C3%A9cnicas-generales/t%C3%A9cnicas-generales-2-tejiendo-con-5-agujas/.

[10] Z. s. HARRIS, «DISTRIBUTIONAL STRUCTURE,» *WORD,* vol. 10, núm. 2-3, pp. 146-162, 1954.

[11] Y. Benjamini i Y. Hochberg, «Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,» *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 57, núm. 1, pp. 289-300, 1995.

[12] O. Levy, Y. Goldberg i I. Dagan, «Improving Distributional Similarity with Lessons Learned from Word Embeddings,» *Transactions of the ACL,* vol. 3, pp. 211-225, 2015.

[13] M. Stamp, Information Security: Principles and Practice, 3rd Edition, 2021.

[14] H. Beker i F. Piper, Cipher Systems: the Protection of Communications, Londres: Northwood Books, 1982.

[15] L. A. B. Sanguino, «Spanish Strip Cipher — Part 1,» MysteryTwister C3 Challenge, Març 2014. [En línia]. Available: https://mysterytwister.org/media/challenges/pdf/mtc3-sanguino-01-SSC-01-en.pdf.

[16] J. L. B. Moreno, «Cryptanalysis of Spanish Civil War Ciphers,» Chair for Embedded Security, Ruhr-Universität Bochum, Bochum, 2013.

[17] M. B. Saz, «Mujeres con ciencia,» 26 Març 2024. [En línia]. Available: https://mujeresconciencia.com/2024/03/26/las-espias-que-tejen-mensajes-secretos-dos-al-derecho-tres-al-reves/.

[18] F. O. Carrero, «Criptografía y resistencia: Códigos secretos en la lucha de las mujeres contra la opresión.,» 2025. [En línia]. Available: https://idus.us.es/items/46b22d38-08f4-47fc-8ac5-05de3234de03.

[19] N. Zarrelli, «The Wartime Spies Who Used Knitting as an Espionage Tool,» 2022 març 16. [En línia]. Available: https://www.atlasobscura.com/articles/knitting-spies-wwi-wwii.

[20] V. Chandola, A. Banerjee i V. Kumar, «Anomaly Detection: A Survey,» 2007. [En línia]. Available: https://www.researchgate.net/publication/220565847_Anomaly_Detection_A_Survey.

ANNEXOS

A.1 Diagrama de Gantt

A continuació es mostra el diagrama de Gantt amb la planificació temporal del projecte:

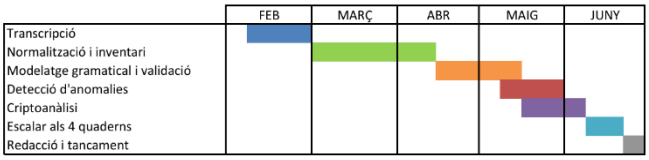


Figura 1. Diagrama de Gantt del TFG, amb les principals fites i períodes de treball.

A.2 Pipeline de dades

A continuació es mostra l'estructura i el flux de les dades en el projecte:

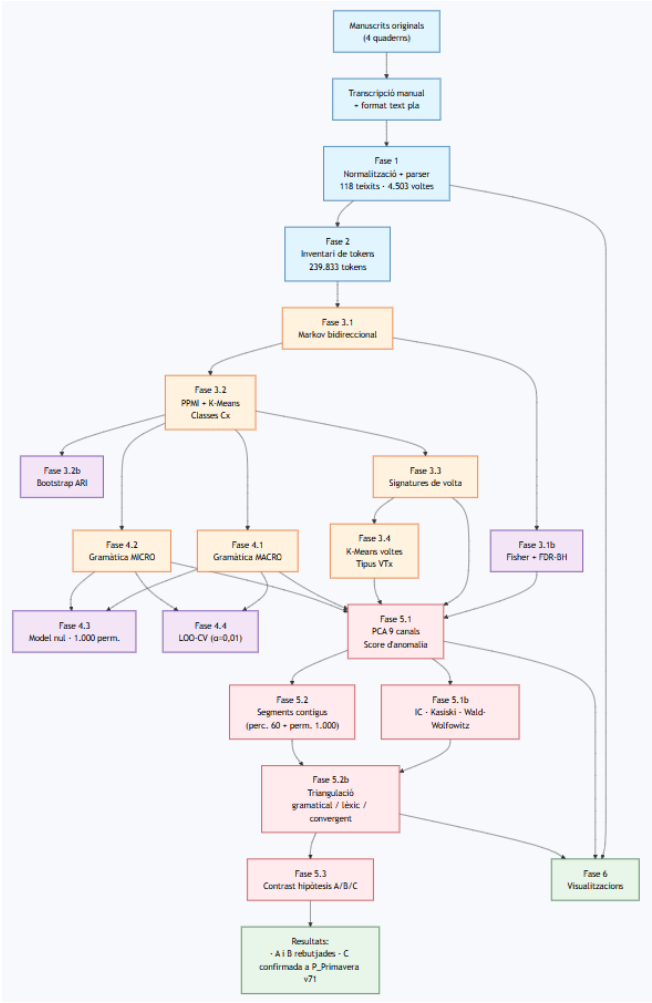


Figura 2. Pipeline de dades (scripts de python)

A.3 Hiperparàmetres

A continuació es mostra els valors definit per cadascun dels hiperparàmetres utilitzat al llarg del pipeline:

Paràmetre	Valor	Script
RANDOM_STATE	42	config.py
PPMI_SMOOTH_ALPHA	0,75	fase3.2
MIN_COUNT_BIGRAM	3	fase3.1b
ALPHA_FDR	0,05	fase3.1b
N_BOOTSTRAP	200	fase3.2b
SMOOTHING_K	0,5	fase4.1, fase4.2
N_PERM (model nul)	1.000	fase4.3
ALPHA_SMOOTH (LOO)	0,01	fase4.4
Z_THRESHOLD	2,0	fase5.1
KASISKI_NGRAM_SIZES	[3, 4, 5]	fase5.1b
RUNS_Z_THRESHOLD	3,0	fase5.1b
THRESHOLD_PERCENTILE	60	fase5.2
W_GRAM, W_LEX	1,0	fase5.2
Hipòtesi B (finestra, llindar)	8, 85%	fase5.3
Hipòtesi A (shuffle ctrl)	15–20	fase5.3

Taula 6. Valors dels hiperparàmetres utilitzats en el pipeline de dades.

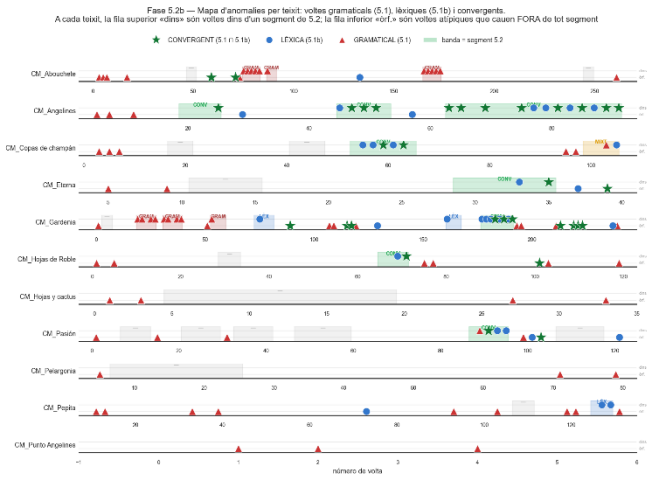
A.4 Resultats anomalies per quadern i teixit

A continuació els segments i voltes anòmals pels teixits del quadern de Negro Demetri.



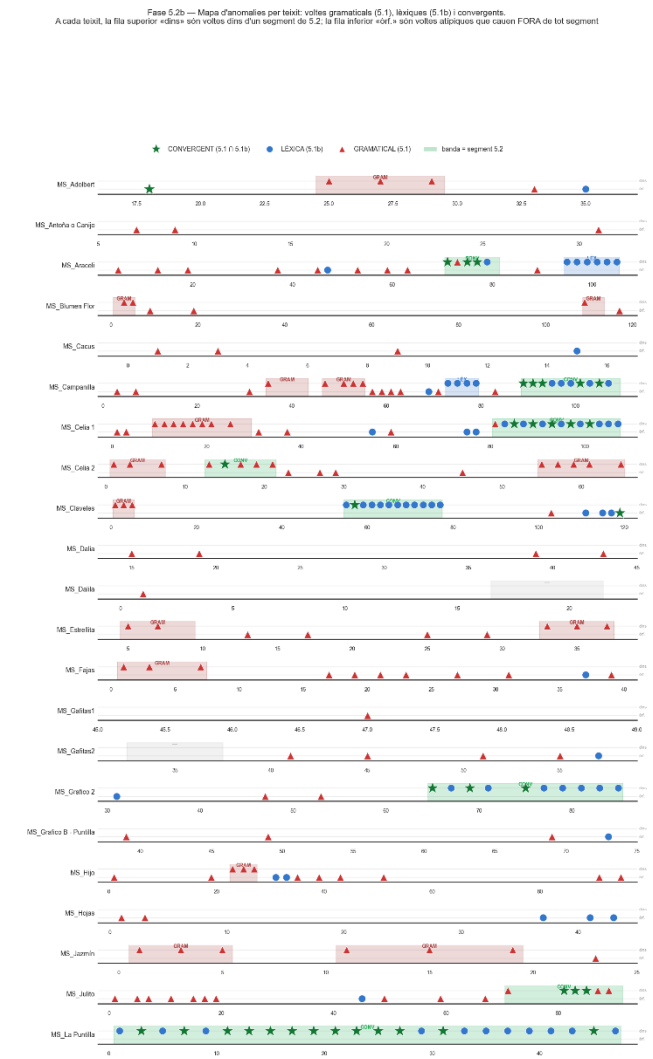
Gràfica 6. Mapa de triangulació gramatical × lèxica per teixit. Les estrelles verdes indiquen voltes convergents; els triangles vermells, perfil gramatical; els cercles blaus, perfil lèxic. (quadern ND)

A continuació els segments i voltes anòmals pels teixits del quadern de Carmen Machado.



Gràfica 7. Mapa de triangulació gramatical × lèxica per teixit. Les estrelles verdes indiquen voltes convergents; els triangles vermells, perfil gramatical; els cercles blaus, perfil lèxic. (quadern CM)

A continuació els segments i voltes anòmals pels teixits del quadern de Manoli Segovia.





Gràfica 8. Mapa de triangulació gramatical × lèxica per teixit. Les estrelles verdes indiquen voltes convergents; els triangles vermells, perfil gramatical; els cercles blaus, perfil lèxic. (quadern MS)

A.5 Mostra d'un teixit transcrit

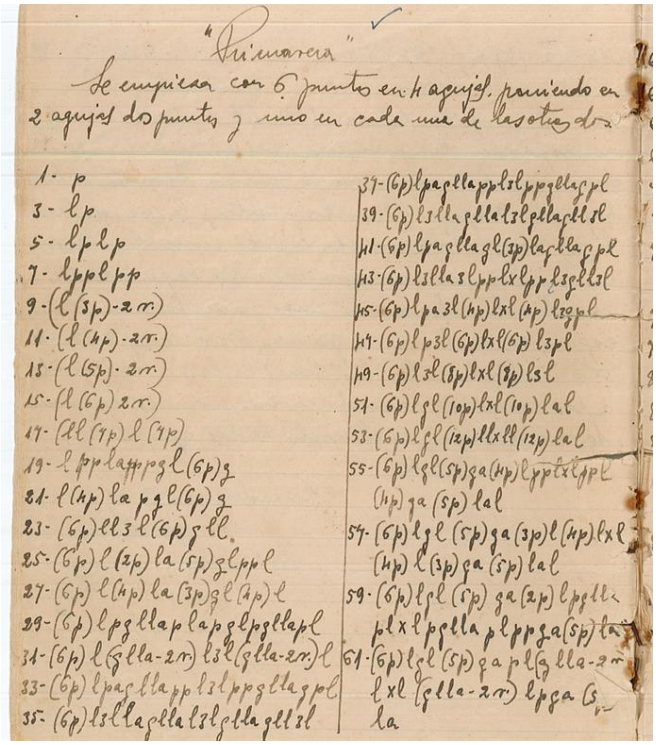


Figura 3. Detall del manuscrit original (P_Primavera)

<u>*Transcripció:*</u>
6p
1 p
3 lp
5 lplp
7 lpplpp
9 (l(3p)-2v)
11 (l(4p)-2v)
13 (l(5p)-2v)
15 (l(6p)-2v)
17 (ll(7p)l(7p)
19 llpplappgl(6p)g
21 l(4p)lappgl(6p)g
23 (6p)ll3l(6p)gll
25 (6p)l(2p)la(5p)glppl
27 (6p)l(4p)la(3p)gl(4p)l
29 (6p)lpgllaplapglpgllapl
31 (6p)l(glla-2v)l3l(glla-2v)l
33 (6p)lpagllappl3lppgllagpl
35 (6p)l3llagllal3lgllagll3l
37 (6p)lpagllappl3lppgllagpl
39 (6p)l3llagllal3lgllagll3l
41 (6p)lpagllagl(3p)lagllagpl
43 (6p)l3llal3lpplxlppl3gll3l
45 (6p)lpa3l(4p)lxl(4p)l3gpl
47 (6p)lp3l(6p)lxl(6p)l3pl
49 (6p)l3l(8p)lxl(8p)lsl
51 (6p)lgl(10p)lxl(10p)lal
53 (6p)lgl(12p)llxll(12p)lal
55 (6p)lgl(5p)ga(4p)lpplxlppl(4p)ga(5p)lal
57 (6p)lgl(5p)ga(3p)l(4p)lxl(4p)l(3p)ga(5p)lal
59 (6p)lgl(5p)ga(2p)lpgllaplxlpgllaplppga(5p)la
61 (6p)lgl(5p)gapl(glla-2v)lxl(glla-2v)lpga(5p)la