Universitat de Lleida

TREBALL FINAL DE GRAU

ESCOLA POLITÈCNICA SUPERIOR
UNIVERSITAT DE LLEIDA
INSPIRING THE FUTURE

Estudiant: Pol Marsol Torras

Titulació: Grau en Enginyeria Informàtica

Títol de Treball Final de Grau: **Neurocognitive Foundations of Oral Discourse Production and Its Predictive Value in Dyslexia-Related Learning Disorders**

Director/a: **Jordi Mateo Fornés**

Presentació

Mes: Juny

Any: 2026

Contents

List of Figures

List of Tables

Abstract

Developmental dyslexia affects approximately 5–10 % of school-age children, yet clinical screening tools remain largely reactive: most instruments are designed for children aged seven and above, after reading failure has already occurred and its academic and psychological consequences have begun to compound. The preschool years of three to six represent a critical window in which neurobiological risk markers manifest as measurable deficits in oral language, offering an opportunity for early identification that current tools fail to exploit.

This work presents NECODYS (NEuroCOgnitive markers of DYSlexia), a prototype early-screening support system for children aged four to five. The system captures spontaneous Catalan oral speech samples through a Flutter mobile application and processes them through a multi-stage linguistic analysis pipeline: automatic transcription, NLP morphosyntactic feature extraction, large language model qualitative profiling, and a discourse quality score on a 0–10 scale that provides the specialist with a quantitative decision-support indicator.

To select and evaluate each pipeline component, a development corpus of manually authored Catalan speech samples representing three discourse quality levels was constructed and benchmarked. The extracted linguistic metrics — including mean utterance length, lexical diversity, and repetition rate — reliably discriminated between quality levels, and the automatic discourse quality score showed strong agreement with researcher-assigned reference scores.

The results demonstrate that end-to-end automated oral discourse screening for Catalan-speaking preschool children is technically feasible. NECODYS is designed as a decision-support tool for speech-language pathologists, not as an autonomous diagnostic system. Clinical deployment requires a prospective study with real patient recordings, independent expert raters, and ethics committee approval.

Acknowledgements

The development of NECODYS has been an enriching journey across a broad spectrum of technologies from speech recognition and natural language processing to large language models, mobile development, and cloud deployment each of which has expanded both the technical depth and the breadth of this work.

I would like to express my sincere gratitude to Dr. Jordi Mateo Fornes, professor and researcher at the Universitat de Lleida, for the invaluable opportunity to contribute to a project originating from the IRB Lleida research group, at the intersection of computing and biomedical research. His guidance throughout the project, and the knowledge gained both through his supervision and in his courses at the Universitat de Lleida, have been fundamental to its completion.

I am also grateful to Joan Maresma for his valuable collaboration throughout this project. His observations, ideas for improvement, and suggestions on how the system could be extended, together with his knowledge of emerging technologies have genuinely enriched both the work and the learning process behind it. The many conversations about computer science, including those held far from campus, have been as instructive as they have been enjoyable.

Chapter 1

Introduction

1.1 Context and Motivation

Developmental dyslexia (hereafter: dyslexia) is a persistent neurodevelopmental disorder that affects approximately 5–10% of school-age children in alphabetic writing systems [1]. Its hallmark is a severe and specific difficulty in acquiring accurate and fluent reading and spelling skills, despite adequate general intelligence, sensory ability, and educational opportunity [2]. Contemporary evidence firmly establishes dyslexia as a biologically grounded condition rooted in atypical maturation of left-hemispheric language networks, not a consequence of reduced intelligence or insufficient instruction [2].

Critically, the neurobiological alterations underlying dyslexia do not first manifest when a child fails to learn to read; they can be detected years earlier, during the preschool period between the ages of three and six. Longitudinal research demonstrates that children later diagnosed with dyslexia exhibit measurable deficits in oral language, phonological awareness, and narrative coherence well before formal literacy instruction begins [3]. The predictive relationship between preschool oral language ability and subsequent reading achievement is robust: children with language impairments at age four carry a substantially elevated risk of reading failure at age eight and beyond [4]. This points to a critical developmental window in which early identification is not only possible but far more effective than post-failure intervention.

Despite this well-established window, clinical practice and available screening tools remain largely reactive. Established dyslexia assessment instruments primarily target children aged seven and above, when reading failure has already occurred and its academic and psychological consequences have begun to compound. No validated automated tool currently exists for systematic screening of children aged three to six years based on spontaneous oral speech. This gap between scientific knowledge and available screening

technology constitutes the primary motivation for the present work.

1.2 Neuroscientific Motivation

Contemporary research identifies the left inferior frontal gyrus (IFG)—broadly corresponding to Broca's area—as a central hub for verbal planning, phonological processing, and the morphosyntactic assembly required for coherent speech production [5, 2]. In pre-literate children, structural and functional anomalies in this region and in its white matter connections to posterior language areas have been documented prior to the onset of formal reading instruction. The phonological deficit hypothesis, the dominant cognitive-level account of dyslexia, establishes that impaired phonological representations disrupt not only grapheme-phoneme decoding but also the higher-order language planning operations required for organised discourse [2]. This suggests a theoretical pathway in which atypical maturation of frontal-inferior language networks would manifest behaviourally as reduced oral discourse complexity, coherence, and morphosyntactic richness in preschool children at risk. This framing constitutes the scientific motivation for NECODYS; the present work does not test this neuroimaging hypothesis directly and makes no causal neurobiological claims.

1.3 The NECODYS System

NECODYS (NEuroCOgnitive markers of DYSlexia) is a prototype early-screening support system designed to capture, transcribe, and analyse oral speech samples from children aged four to five years. The system comprises a Flutter mobile application that guides a specialist through a structured evaluation session, a cloud backend that transcribes recordings using a state-of-the-art Catalan speech recognition model, a natural language processing (NLP) pipeline that extracts quantitative linguistic indicators from the transcription, a large language model (LLM) component that performs morphosyntactic profiling, and a discourse quality score (DQS) on a 0–10 scale that provides the specialist with a calibrated decision-support indicator. NECODYS is designed as a decision-support tool for speech-language pathologists and clinical specialists, not as an autonomous diagnostic system.

This work focuses on technical feasibility and does not involve the collection of data from real children. The development corpus used for system calibration and benchmarking consists of manually authored sentences representing three discourse quality levels. Any future clinical deployment would require ethics committee approval and prospective validation with real patient populations.

1.4 Research Objectives

The following objectives were pursued in this work:

O1. Develop a mobile application and backend system for recording and analysing oral speech samples from 4–5 year-old children, extracting morphosyntactic linguistic indicators.

O2. Implement and evaluate a multi-stage linguistic analysis system, demonstrating that extracted NLP and LLM metrics significantly discriminate between discourse quality levels.

O3. Design and calibrate a discourse quality score (0–10) showing significant agreement with expert reference annotations, providing specialists with a quantitative decision-support indicator.

O4. Evaluate the technical feasibility of the proposed system as an early-screening prototype for linguistic risk indicators associated with dyslexia.

O5. Design and implement a mobile application capable of guiding specialists through structured evaluation sessions, collecting oral language samples, and integrating automated linguistic analysis into a single workflow.

Chapter 2

Related Work

2.1 Neuroscientific Foundations of Developmental Dyslexia

The dominant cognitive account of developmental dyslexia is the phonological deficit hypothesis (PDH), which holds that impaired phonological representations disrupt the mapping of graphemes to phonemes required for reading acquisition [5, 2]. An influential extension, the double-deficit hypothesis, further argues that independent deficits in phonological processing and in rapid automatised naming (RAN) can each impair reading fluency, and that their co-occurrence produces the most severe outcomes [6]. Epidemiological studies estimate the prevalence of developmental dyslexia at 5–10% of school-age children in alphabetic orthographies [7].

Neuroimaging research has established that these cognitive deficits have a clear biological substrate. Structural and functional anomalies in the left-hemisphere reading network—particularly in the left inferior frontal gyrus (IFG), temporo-parietal cortex, and occipito-temporal regions—have been documented in pre-literate children prior to the onset of formal reading instruction [8]. Longitudinal studies using multivariate fMRI further demonstrate that neural activation patterns in these regions predict future reading outcomes with high accuracy [9]. White matter tractography confirms reduced microstructural integrity of the arcuate fasciculus, the pathway linking frontal and posterior language regions, in children at familial risk [10]. Compensatory activation of the right IFG has been observed in children who eventually reach adequate reading levels, suggesting neuroplastic reorganisation over development [11].

Beyond the core phonological account, complementary theories implicate auditory temporal processing [12] and cerebellar motor timing [13]. While the relative weight of these accounts remains debated, the phonological deficit hypothesis retains the most direct clinical implications: it explains why oral language skills acquired years before formal reading instruction

serve as reliable early markers of risk.

2.2 Early Oral Language Markers and the Preschool Window

Longitudinal evidence robustly establishes that the neurobiological alterations underlying dyslexia manifest in oral language long before a child first encounters written text. Scarborough demonstrated that children later diagnosed with dyslexia exhibit measurable deficits in phonological awareness, vocabulary, and narrative coherence as early as age two and a half [3]. A large prospective cohort study confirmed that children with language impairment at age four carry a substantially elevated risk of reading failure at age eight and beyond [4]. Despite this well-established predictive relationship, most standardised screening instruments target children aged seven and above, after reading failure has already occurred [1].

A clinically important complication in preschool assessment is the overlap between developmental dyslexia and Developmental Language Disorder (DLD). Bishop and Snowling conceptualised this relationship through a two-dimensional framework: dyslexia is primarily localised to a phonological processing deficit, while DLD encompasses a broader disruption of morphosyntactic and semantic networks [14]. The Phase 2 consensus statement on DLD terminology formalised this distinction and highlighted the need for differential oral language profiling in early assessment [15]. For the purposes of preschool screening, the practical consequence is the same: children at risk for either condition exhibit reduced morphosyntactic richness and narrative coherence in spontaneous speech.

Spontaneous narrative production has emerged as an ecologically valid assessment paradigm precisely because it imposes simultaneous demands on working memory, lexical retrieval, and syntactic planning [16]. When these resources are taxed together, underlying fragilities surface in measurable ways: reduced mean length of utterance (MLU), lower type-token ratio (TTR), shallower syntactic dependency depth, fewer connectors, and elevated repetition rates. Automated extraction of these metrics from transcribed speech samples constitutes the computational core of modern preliteracy screening.

2.3 Existing Digital Screening and Intervention Systems

The landscape of digital tools for dyslexia has expanded substantially over the past decade. A systematic review identified eight platforms, games, and videogames with published empirical evidence marketed in Spain and anal-

ysed them across neurological, cognitive, psycholinguistic, and performance levels of intervention [17]. Table 2.1 summarises the systems identified in the literature, separating screening instruments from intervention programmes.

The table reveals a consistent structural gap: the only tools validated for pre-reading children (EarlyBird, Berni) rely on structured tasks with controlled stimuli, not on spontaneous oral speech; all tools requiring baseline reading ability are inaccessible to the preschool population; and none operates in Catalan. NECODYS targets all three of these gaps as a decision-support prototype.

2.4 NLP and Automated Speech Analysis in Clinical Assessment

Automated NLP pipelines enable the objective extraction of the developmental linguistic markers discussed in Section 2.2. Table 2.2 lists the five core metrics targeted by NECODYS, together with their computational basis and clinical interpretation [16].

Recent work on artificial intelligence in neurodevelopmental assessment highlights the potential of computational methods while acknowledging their limitations [27]. Syntactic awareness in early childhood has been identified as a particularly robust predictor of later reading ability, reinforcing the case for automated morphosyntactic profiling in the preschool period [28]. Nevertheless, two critical gaps remain unaddressed: no validated ASR pipeline targeting Catalan child speech in the preschool age range has been reported in the literature, and the application of LLMs to morphosyntactic profiling for decision support in pre-literacy assessment is novel in this demographic. NECODYS contributes a first end-to-end prototype implementation—Catalan ASR, NLP metric extraction, LLM morphosyntactic analysis, and a calibrated discourse quality score (DQS) on a 0–10 scale—evaluated against expert annotations on a three-level development corpus.

Tool	Age	Type	Method & Evidence	Key Limitation
EarlyBird	4–7 years	Screening	Computer Adaptive Testing with Item Response Theory scoring (CAT/IRT) of oral language and phonological awareness; AUC = 0.88, 85% correct classification [18]	No spontaneous speech analysis
Berni	4–6 years	Screening	Autonomous play-based; significant efficacy vs. control group [19]	Geographically limited (Basque Country)
Dytective	7+ years	Screening	Random Forest on HCI data; ≈83% accuracy (Spanish cohort); sensitivity prioritised over specificity to minimise false negatives [20]	Requires reading ability
Lexplore	6+ years	Screening	AI-driven analysis of eye-tracking data (saccades, fixations) during reading; identifies reading disabilities with statistically significant group separation [21]	Specialised hardware; reading required
SICOLE-R	7–12 years	Screening	Multimedia cognitive battery; validated discriminant validity [22]	Reading-level dependent
UBinding	6–7 years	Intervention	Blended school+home phonics; word-reading $d = 0.99$ [23]	Intervention only, not screening
Galexia	8+ years	Intervention	Repeated reading on Android; significant reading fluency gains [24]	Targets post-acquisition fluency
GraphoGame	5–9 years	Intervention	Adaptive phonics; significant phonological decoding gains (Hedges' $g = 0.36$) in disadvantaged readers [25]	No discourse analysis

Notes. AUC: area under the receiver operating characteristic curve (1.0 = perfect discrimination; chance = 0.5). Cohen's d and Hedges' g: standardised mean-difference effect sizes quantifying practical magnitude of a treatment or group difference ($d/g \approx 0.2$ small, 0.5 medium, 0.8 large per Cohen's conventions [26]).

Table 2.1: Existing digital systems for dyslexia screening and intervention.

Metric	**NLP Extraction**	**Clinical Significance**
MLU	Mean words per utterance (sentence segmenter)	Expressive language level; reduced in Specific Language Impairment (SLI) and dyslexia
TTR	Unique tokens / total tokens	Lexical diversity; low TTR indicates word-retrieval deficit
Syntactic depth	Max and mean dependency tree depth (parser)	Complex clause generation; proxy for working memory capacity
Connector rate	POS-tagged conjunctions and discourse markers	Discourse cohesion; impaired in children with narrative difficulties
Repetition rate	Consecutive near-identical token sequences	Phonological planning load; primary dyslexia risk marker

Table 2.2: Core NLP metrics for automated oral discourse assessment.

Chapter 3

Methods

3.1 Development Corpus

A development corpus of 130 manually authored Catalan samples was constructed for system calibration and comparative benchmarking across all three pipeline stages. It comprises 120 child-speech samples distributed equally across three discourse quality levels—Level A (high dyslexia-risk profile, $n = 40$), Level B (intermediate profile, $n = 40$), and Level C (good discourse level, $n = 40$)—plus ten adult-authored complex sentences included to verify the upper boundary of the lexical richness scale.

Level assignment followed established clinical characterisation criteria for preschool oral language profiles [16, 15]. Level A utterances are telegraphic (1–5 words), contain at least one consecutive word repetition, and show no connectors or subordination. Level B utterances are transitional (5–10 words), feature occasional repetitions, and employ simple paratactic coordination. Level C utterances are structured (8 or more words) and include embedded subordination (temporal, causal, or relative clauses), lexically varied vocabulary, and connectors. Level assignments were made by the researcher; independent inter-rater validation of corpus quality levels is an acknowledged methodological limitation and is required prior to any clinical deployment.

No real children participated in this study; all samples were manually authored to represent target clinical profiles. Results obtained on this corpus constitute system calibration evidence, not validated clinical performance. Any future clinical deployment requires ethics committee approval and a prospective study with real patient data.

3.2 Model Selection Methodology

Each pipeline stage was evaluated against the development corpus using stage-appropriate criteria. The subsections below describe the evaluation

protocol and selection rationale for the speech-to-text (STT), natural language processing (NLP), and large language model (LLM) components.

3.2.1 Speech-to-Text Evaluation

Catalan-specific ASR models were investigated as additional candidates, including:

- `projecte-aina/whisper-large-v3-ca-3catparla`
- `BSC-LT/whisper-bsc-large-v3-cat`
- `softcatala/wav2vec2-large-xlsr-catala`

These models could not be included in the benchmark for two compounded reasons: none had the HuggingFace Serverless Inference API activated at the time of evaluation (all requests returned `HTTP 503---modelnotloaded`), and running them locally requires dedicated GPU hardware that was not available in this project. The STT benchmark is therefore restricted to Whisper-architecture models accessible via Groq cloud inference (H100 GPU) or local CPU-based inference.

Six STT candidates based on the Whisper architecture [29] were evaluated on all 120 audio samples (the ten adult corpus samples are text-only and carry no associated audio recordings; 720 total inferences): `whisper-large-v3` and `whisper-large-v3-turbo` via Groq cloud inference (H100 GPU), and four local `faster-whisper` variants (tiny, base, small, medium). Four metrics were computed:

- **WER** (word error rate): $(S+I+D)/N_{\text{ref}}$, where S, I, D are word-level substitutions, insertions, and deletions.
- **CER** (character error rate): the same measure computed at character level; sensitive to partial phonological errors that WER counts as full-word misses.
- **RTF** (real-time factor): processing time divided by audio duration; RTF < 1 is required for real-time mobile deployment.
- **DPR** (disfluency preservation rate): proportion of consecutive word repetitions present in the reference transcription that are preserved in the model output (DPR = preserved repetitions/total reference repetitions). Computed on the 84 of 120 samples containing at least one repetition. DPR is clinically critical because consecutive repetitions are the primary acoustic disfluency marker in high-risk child speech.

Table 3.1 reports the benchmark results for all six candidates.

Model	WER	CER	RTF	DPR_A (%)
`whisper-large-v3` (Groq)	0.176	0.075	0.47	67.2
`faster-whisper-medium`	0.210	0.083	100.5	64.3
`faster-whisper-small`	0.229	0.096	28.0	53.6
`whisper-large-v3-turbo` (Groq)	0.274	0.126	0.56	38.7
`faster-whisper-base`	0.357	0.154	10.7	51.8
`faster-whisper-tiny`	0.398	0.172	0.18	50.6

Table 3.1: STT benchmark results ($n = 120$, sorted by WER). DPR_A: disfluency preservation rate on Level A samples (84 samples containing repetitions). Lower WER, CER, and RTF are better; higher DPR_A is better. `faster-whisper-medium` is operationally excluded by its RTF of 100.5× real-time despite its competitive DPR.

Among the six Whisper-architecture candidates, `whisper-large-v3` is the only model that simultaneously satisfies all three operational thresholds: faster-than-real-time processing (RTF < 1), a word error rate below 0.20, and a disfluency-preservation rate above 60% on the high-risk Level A samples (DPR_A).

To verify that this lower error rate reflects a genuine quality difference rather than chance variation across the 120 audio samples, `whisper-large-v3` was compared against each of the other five models with a Wilcoxon signed-rank test [30]. Intuitively, this test sorts the per-sample differences in error between two models and asks whether they lean consistently in one direction; because it works on the *ranks* of those differences rather than their raw magnitudes, it does not assume the errors follow a normal (bell-shaped) distribution. This makes it more appropriate here than the more familiar paired t-test, since the WER values are strongly right-skewed (a few samples have very high error) and would violate that normality assumption. Running five comparisons at once raises the chance that at least one would appear significant by accident, so the significance threshold was tightened with a Bonferroni correction [31]: the conventional level of $\alpha = 0.05$ was divided by the five tests, yielding a stricter adjusted threshold of $\alpha_{\text{adj}} = 0.01$. Under this criterion, `whisper-large-v3` significantly outperformed `faster-whisper-tiny` ($p < 0.001$), `faster-whisper-base` ($p < 0.001$), and `whisper-large-v3-turbo` ($p = 0.0005$).

`faster-whisper-small` was marginal ($p = 0.037$) and did not survive Bonferroni correction. `faster-whisper-medium` was not significantly different ($p = 0.268$) but is operationally excluded by its RTF of 100.5× real-time. Figure 3.1 illustrates the quality–latency trade-off across all candidates.

One systematic limitation requires explicit acknowledgement: Whisper is trained on clean adult speech and exhibits a normalisation bias that silently removes approximately 33% of consecutive word repetitions present in child utterances. This directly reduces DPR and may undercount the primary

clinical disfluency marker in Level A transcriptions; this limitation is discussed further in Chapter 6.

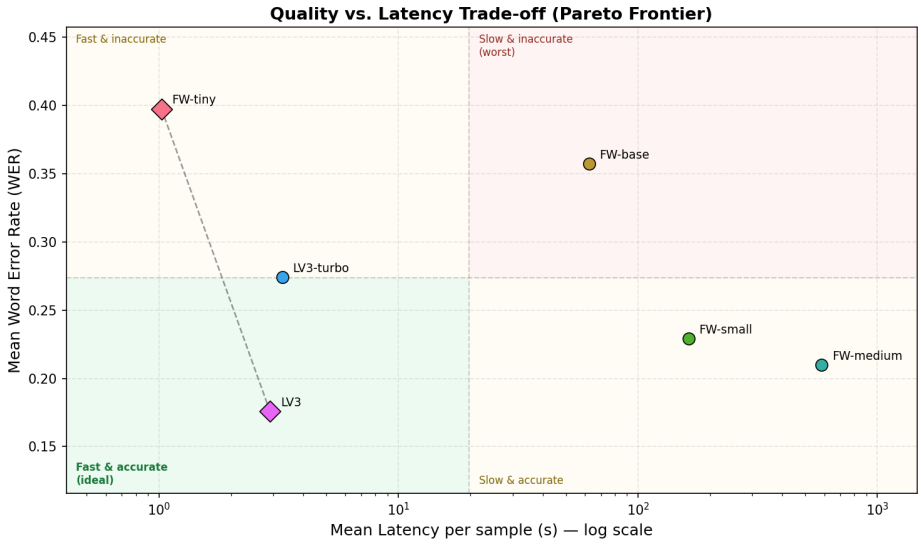


Figure 3.1: Model selection for the STT component. Pareto scatter of word error rate (WER) versus real-time factor (RTF, log scale) for six candidate models ($n = 120$). The viable deployment region requires RTF < 1 (faster than real-time) and low WER; only `whisper-large-v3` satisfies both constraints simultaneously.

3.2.2 NLP Metric Validation

Morphosyntactic features were extracted using spaCy [32] with the `ca_core_news_sm` model (v3.8.0, trained on Catalan news text). The transformer variant (`ca_core_news_trf`, ≈500 MB) was also tested; it degraded accuracy specifically at Level A — the high-risk group most critical for screening — and was rejected in favour of the lighter model.

To establish which metrics are reliable for downstream analysis, all 130 corpus samples were manually annotated for five binary features (complex sentence, subject present, direct object present, has repetitions, predominant verbal tense) and connector frequency. Agreement between manual annotations and automatic spaCy output was measured using Cohen κ for binary features and Spearman ρ for connector count. Cohen's κ quantifies agreement corrected for chance: it ranges from 0 (agreement no better than random) to 1 (perfect agreement), with conventional interpretation thresholds of $\kappa < 0.20$ (slight), 0.20–0.40 (fair), 0.40–0.60 (moderate), 0.60–0.80 (substantial), and > 0.80 (almost perfect) [26]. For example, a specialist and the automatic system agreeing on "has repetitions" by pure chance (50 %

base rate) would give $\kappa = 0$; perfect agreement on every sample gives $\kappa = 1$. Figure 3.2 shows the per-level agreement breakdown.

Agreement ranged from near-chance to good: predominant verbal tense ($\kappa = 0.796$), direct object ($\kappa = 0.632$), repetition detection ($\kappa = 0.494$), subject presence ($\kappa = 0.380$), and complex sentence ($\kappa = 0.014$). The last result reflects an architectural limitation: spaCy captures subordination via dependency relations but does not detect coordination (e.g., *menja i dorm*), causing systematic false negatives for this feature. Connector count showed strong ordinal agreement ($\rho = 0.916$, $p < 0.001$) but absolute values are overestimated by approximately 2×, as consecutive coordinator repetitions are each tagged as connectors by the POS tagger. Verb count is unreliable at Level A, where 50% of fragmented utterances receive zero verb predictions.

The metrics retained as primary reliable indicators are: mean length of utterance (MLU), as defined in Section 2.4, type-token ratio (TTR), repetition rate, syntactic tree depth, subordination index, and predominant verbal tense. The raw complex-sentence count is replaced by the subordination index (fraction of utterances containing a syntactic subordinate clause), which is more precisely defined and aligns better with manual annotation. Connector count and verb count are retained as secondary features with documented caveats.

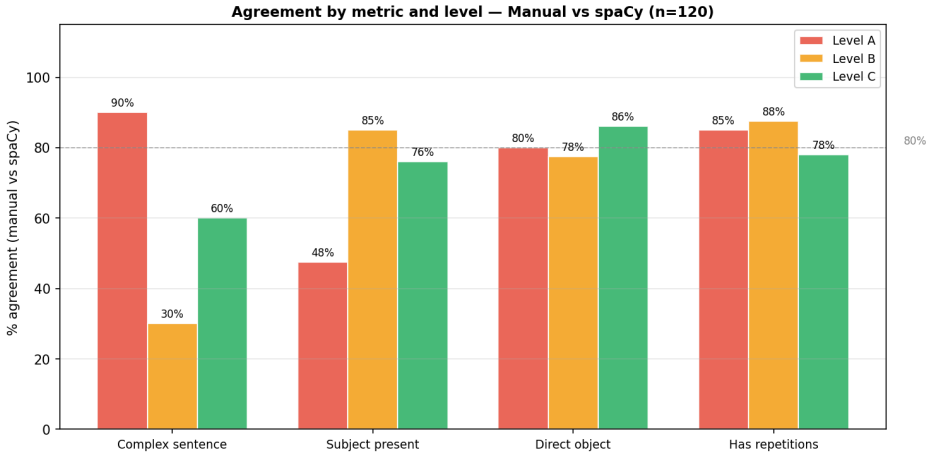


Figure 3.2: Agreement between automatic spaCy annotations and manual reference annotations across four binary metrics and three discourse quality levels ($n = 130$). The dashed line at 80% indicates the author-defined reliability threshold: metrics above this level are retained as primary reliable indicators with no caveats; metrics below it are documented as having a known limitation but may be retained as secondary features if they remain statistically significant across quality levels. Complex sentence detection falls systematically below threshold owing to the architectural limitation that dependency parsing captures subordination but not coordination.

3.2.3 LLM Morphosyntactic Analysis

Two LLM candidates were fairly evaluated on the morphosyntactic analysis task: `gemini-3.1-flash-lite-preview` (Google DeepMind) and `llama-3.1-8b-instant` (Meta, via Groq), yielding 240 inferences in total (2×120 child samples; the ten adult samples are excluded from this comparative benchmark to avoid artificially inflating discrimination metrics, since all ten are correctly classified as high lexical richness by design). A third candidate, `llama-3.3-70b-versatile` (70B parameters, Groq), was attempted but is excluded from this comparison: 41 of 120 inference calls failed due to API rate limiting, making reliable metric computation impossible. Additionally, three Gemini models could not be evaluated due to exhaustion of the free-tier daily quota. Both exclusions are documented as limitations in Chapter 7.

Each model received the transcribed text and was instructed to return a structured JSON report covering four sections: quantitative metrics, morphosyntactic diagnosis, lexical-semantic analysis, and pragmatic analysis. Two criteria were applied. First, *schema compliance*: the proportion of the 120 responses returning a syntactically valid JSON object with all required fields present — a hard production requirement, since non-compliant responses cannot be processed by downstream pipeline components. Second, *clinical discrimination*: Spearman ρ between the model-assigned lexical richness category (low = 1, medium = 2, high = 3) and the ground-truth discourse quality level (A = 1, B = 2, C = 3).

`gemini-3.1-flash-lite-preview` was selected as the only model among the two fairly evaluated candidates achieving both 100% schema compliance and $\rho > 0.5$ simultaneously: `llama-3.1-8b-instant` reaches 95.8% schema compliance but yields $\rho = 0.386$, below the $\rho \geq 0.5$ informativeness threshold (a researcher-defined minimum: $\rho = 0.5$ marks the boundary between weak and moderate rank correlation; below this level a model cannot meaningfully separate the three discourse quality groups in clinical practice, making it unsuitable as a selection criterion [4]).

`gemini-3.1-flash-lite-preview` was selected as the LLM component. One distributional characteristic of its predictions merits documentation: lexical richness was compressed to two categories (low and medium) for all 120 child-speech samples, and the high category was never assigned. This reflects the inherent complexity ceiling of the development corpus rather than a model defect: when applied to the ten adult-authored complex sentences, the model correctly classified all ten as high lexical richness. This phenomenon does not affect the relative discrimination utility of the model within the child population but will require re-calibration when clinical samples become available.

3.3 Discourse Quality Score Calibration

The discourse quality score (DQS), a 0–10 scalar indicator produced automatically by the analysis pipeline, was calibrated against expert reference annotations. At the end of each evaluation session the supervising specialist assigns a holistic discourse level score on a continuous 0–10 slider, based on direct observation of the child responses. This score serves as the clinical calibration reference.

A critical design constraint ensures the independence of the two measurements: the expert reference score is never provided to any component of the automated pipeline. The DQS is computed exclusively from the transcription, NLP metrics, and LLM morphosyntactic output. This independence is a prerequisite for interpreting their agreement as a meaningful experimental result rather than a circular artefact.

Agreement between the automatic DQS and the expert reference score is quantified using Bland–Altman analysis [33], which characterises the systematic bias (mean of the differences) and the limits of agreement (mean $\pm 1.96 \cdot \mathrm{SD}$ of the differences) between two measurement methods. Calibration on the development corpus yielded approximate score ranges of DQS ≈ 2.0 for Level A, 4.5–6.5 for Level B, and 6.5–9.5 for Level C. These constitute preliminary calibration evidence; clinical cut-off thresholds require validation with real patient populations. Results are reported in Chapter 6.

The expert reference scores used for calibration were provided by the researcher; independent validation with annotators holding established clinical credentials in speech-language pathology is required before the DQS can be applied in a clinical context.

The selected components and evaluation criteria inform the system architecture described in Chapter 4.

Chapter 4

Architecture

4.1 System Overview

The NECODYS system comprises six components connected by HTTP and asynchronous data flows (Figure 4.1). The mobile client, implemented in Flutter (Dart), guides the specialist through eight structured evaluation tasks, records audio responses via the device microphone, and renders analysis results; Catalan text-to-speech instructions are synthesised locally using the `flutter_tts` package with locale `ca-ES`.

The backend is a FastAPI application deployed as a single-service Docker container on a dedicated Linux server at 192.168.101.74, port 8000. The container runs one Uvicorn worker, mounts two persistent volumes (`storage/` for audio files and `static/` for task images), and loads configuration from a `.env` file via the Compose `env_file` directive. Two cloud inference endpoints handle the computationally intensive tasks: the Groq API (NVIDIA H100 accelerated) performs speech-to-text transcription, and the Gemini API hosts both large language model components.

MongoDB persists application data across three collections: `activities` (audio metadata and transcriptions), `patients` (demographic and clinical records), and `pipeline_results` (analysis outputs). Authentication is handled by Firebase, which issues JSON Web Tokens to the mobile client; the backend verifies each token without storing credentials locally.

4.2 REST Interface and Analysis Pipeline

Table 4.1 lists the REST endpoints exposed by the backend. Every request must carry a Firebase Bearer token in the `Authorization` header; the `get_current_user` FastAPI dependency verifies the token and rejects unauthenticated calls with HTTP 401. Activity upload (`POST /activities/`) accepts a `multipart/form-data` request containing the audio file, the task key, and the patient identifier; transcription is triggered synchronously be-

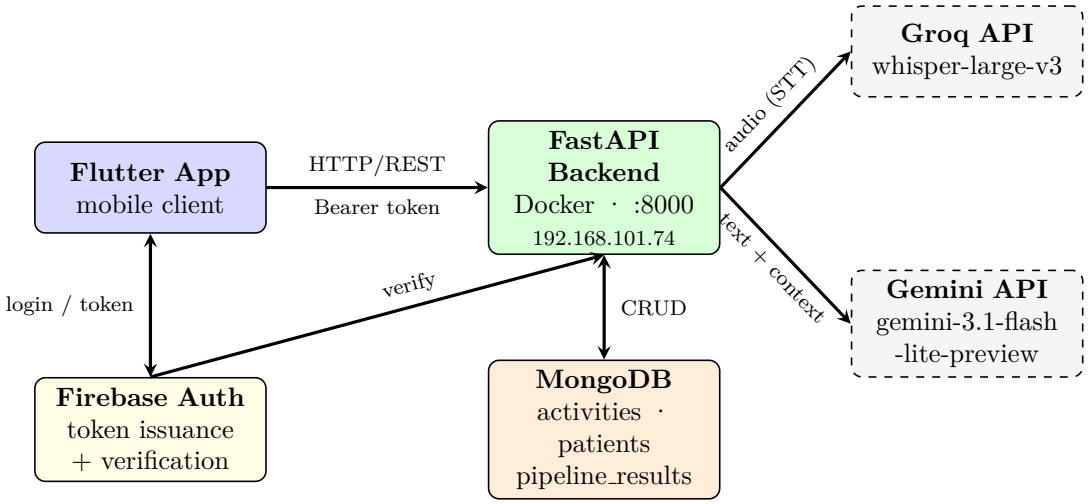


Figure 4.1: NECODYS system architecture. Solid arrows indicate primary data flow; dashed boxes denote external cloud services. The FastAPI backend runs as a Docker container at 192.168.101.74:8000 and serves as the sole integration point between the mobile client, the analysis pipeline, and all external APIs.

fore the Activity document is persisted. Analysis is decoupled from upload: the client explicitly triggers the pipeline via the `/analysis/` router.

Figure 4.2 illustrates the per-task pipeline execution. The handler first queries `pipeline_results` for an existing document with `estat=completed`; if found, the cached result is returned immediately, preventing redundant API calls. On a cache miss, three operations execute concurrently via `asyncio.gather`: NLP feature extraction (spaCy, CPU-bound, wrapped in `asyncio.to_thread`), task definition retrieval, and patient context retrieval from MongoDB. The two LLM calls are sequential, as the morphosyntactic analyser must complete before the discourse quality score (DQS) model receives its output. Any component failure returns an empty dataclass rather than raising an exception, ensuring graceful degradation to a `partial` result. The global session pipeline (`POST /analysis/session/`) iterates over all patient activities sequentially and aggregates per-task DQS scores into a session-level mean.

Figure 4.3 complements the execution flow above by detailing the data transformations at each pipeline stage: what each module receives as input and what structured artefact it produces as output.

Method	Endpoint	Description
POST	`/activities/`	Upload audio file; trigger STT transcription; persist `Activity` document
GET	`/activities/patient/{id}`	List all recordings for a given patient
POST	`/analysis/{id_activitat}`	Execute per-task analysis pipeline; return `PipelineResult` (HTTP 201)
GET	`/analysis/{id_activitat}`	Retrieve latest per-task `PipelineResult`
POST	`/analysis/session/{id_usuario}`	Execute global session pipeline over all patient activities (HTTP 201)
GET	`/analysis/session/{id_usuario}`	Retrieve latest global `PipelineResult`
POST/GET/PATCH	`/patients/`	Patient record management (CRUD)
GET	`/tasks/`	Task definition catalogue

Table 4.1: REST API endpoint summary. All endpoints are authenticated via Firebase (Bearer token). Analysis endpoints are idempotent: a cached `completed` result is returned without re-invoking any external API.

4.3 Data Models

All analysis outputs are encapsulated in a single `PipelineResult` document (Table 4.2), persisted to the `pipeline_results` MongoDB collection and returned verbatim to the mobile client. Each document aggregates the outputs of all pipeline stages — NLP metrics, LLM morphosyntactic analysis, discourse quality score, and clinical summary — into a single structure, with optional fields set to `None` on component failure.

Chapter 5 details the implementation of each architectural component.

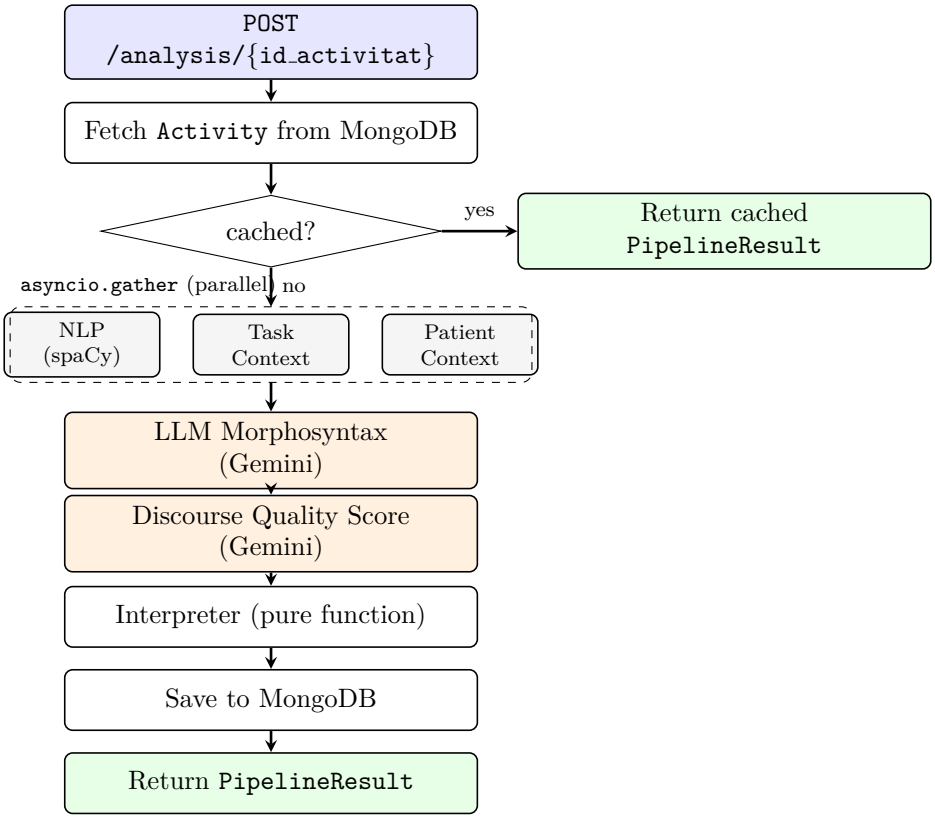


Figure 4.2: Per-task analysis pipeline execution flow (`POST /analysis/{id_activitat}`). Parallel branches inside `asyncio.gather` feed the two sequential LLM steps; component failures degrade gracefully without propagating exceptions.

Field	Type	Description
`id_activitat`	`str`	Activity ObjectId (per-task) or `global_{id}_{ts}` (session)
`nlp_metrics`	`dict?`	20+ spaCy features: MLU, TTR, repetition rate, syntactic tree depth, subordination index, verbal tense, POS counts
`analisi_llm`	`dict?`	Gemini morphosyntactic report: quantitative metrics, morphosyntactic diagnosis, lexical–semantic analysis, pragmatic analysis
`termometre`	`dict?`	DQS [0–10], qualitative label (5 levels), justification text, four sub-indicators
`informe_clinic`	`str?`	Structured clinical summary generated by the interpreter
`estat`	`str`	Pipeline outcome: `completed` / `partial` / `error`

Table 4.2: `PipelineResult` document schema. Optional fields (`dict?`, `str?`) are set to `None` on component failure; `estat` reflects the completeness of the result: `completed` if both NLP and LLM succeeded, `partial` if only NLP succeeded, `error` if NLP failed.

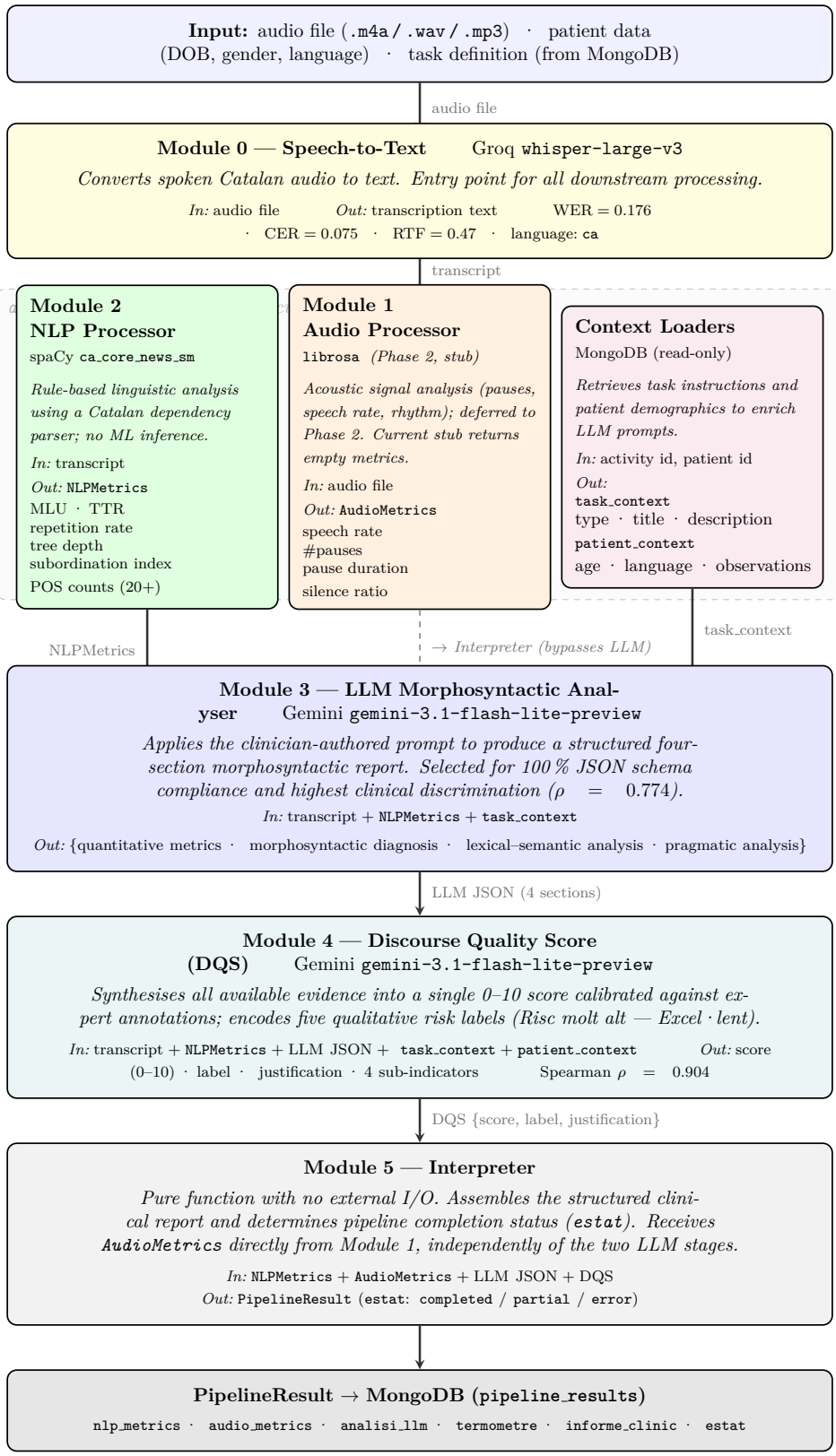


Figure 4.3: Module-level data flow of the analysis pipeline. Each box lists the primary inputs (*In*) and outputs (*Out*) for that stage, together with a brief description of the module's role. The three modules inside the dashed band execute concurrently via `asyncio.gather`, and two data paths bypass the LLM stages (described in the box text above). The Audio Processor is a functional stub pending Phase 2 implementation.

Chapter 5

Implementation

5.1 Speech-to-Text Module

The speech-to-text (STT) component is implemented in `src/services/ai.py` and constitutes the first stage of the analysis pipeline. `whisper-large-v3` via the Groq cloud API was selected as the primary transcription model on the basis of word error rate, real-time factor, and disfluency preservation criteria (Chapter 3); the language is fixed to Catalan to suppress auto-detection latency.

The public entry point `transcribe_audio()` wraps the Groq call in a two-level fallback chain: a Groq exception activates the local `faster-whisper` model with voice-activity detection filtering, ensuring service continuity during API outages. The local model is loaded lazily on first use and managed as a thread-safe singleton. Configuration, including the API key, is loaded from the `.env` file through `pydantic-settings`.

5.2 NLP Analysis Module

Morphosyntactic feature extraction is implemented in `src/pipeline/nlp_processor.py`. The module loads spaCy's `ca_core_news_sm` model lazily on the first request, with `spacy.blank("ca")` as a degraded tokenisation fallback. The public function `process(text)` returns an `NLPMetrics` dataclass containing 25 optional fields organised in five categories (Table 5.1).

Each metric category is computed inside an independent `try/except` block, so a failure in one category sets only the affected fields to `None` without interrupting the remaining categories. Sentence-level dependency features are extracted per utterance and then aggregated into the global `NLPMetrics` instance.

Category	Key Fields	Description
Basic	`lpm`, `ttr`, `total_enunciats`, `index_subordinacio`	Utterance-level statistics. `lpm` (mean utterance length) is the primary discriminating metric; `index_subordinacio` = complex / total utterances.
POS counts	`num_verbs`, `num_connectors`, `num_noms`	Part-of-speech token frequencies. Connector count is systematically overestimated by $\approx 2\times$ owing to consecutive coordinator repetitions.
Syntax	`frases_amb_subjecte`, `profunditat_arbre`	Dependency-tree depth and grammatical role presence per utterance.
Verbal morphology	`temps_present`, `temps_passat`, `temps_predominant`	Verbal tense distribution from spaCy morphological attributes (Cohen $\kappa = 0.796$; Chapter 3).
Repetitions	`ratio_repeticions`, `paraules_repetides`	Ratio of consecutive repeated words to total tokens; primary risk discriminator across discourse quality levels (Chapter 6).

Table 5.1: `NLPMetrics` field categories. All fields are typed `Optional`; `None` is returned if a metric cannot be computed.

5.3 LLM Morphosyntactic Analysis

LLM morphosyntactic profiling is implemented in `src/pipeline/llm_analyzer.py`, using the `google.genai` asynchronous client with structured JSON output enforced at inference level. The model is `gemini-3.1-flash-lite-preview`.

The prompt pre-injects the five quantitative NLP metrics already computed by the preceding stage (`lpm`, `ttr`, `total_enunciats`, `frases_simples`, `frases_complexes`), so the model concentrates its effort on the qualitative sections: morphosyntactic diagnosis, lexical-semantic analysis, and pragmatic analysis. When task context is available (type, title, and instruction given to the child), it is appended to the prompt to calibrate expectations for different task genres. Output is validated against the four-section schema; any failure returns `None` and degrades the pipeline result to `estat = "partial"`.

5.4 Discourse Quality Score

The discourse quality score (DQS) module is implemented in `src/pipeline/llm_termometer.py`. The same Gemini model is invoked in a second sequential call that receives the complete morphosyntactic report alongside the raw transcription and NLP metrics.

The prompt embeds five calibration examples, one per quality level, each pairing a reference transcription, NLP metrics, and morphosyntactic summary with the expected output. Patient context (child age, habitual language, and specialist observations) is prepended before inference.

A critical design constraint governs context injection: the expert clinical score assigned via the application's 0–10 slider (`nivell_discurs`) is **never** provided to any LLM prompt. This independence is a prerequisite for interpreting the agreement between automatic and expert scores—quantified via Bland–Altman analysis in Chapter 6—as a meaningful experimental finding. At the free-tier limit of 15 requests per minute, two Gemini calls per analysis reduce maximum throughput to approximately seven pipeline runs per minute.

5.5 Mobile Application

The mobile client is implemented in Flutter 3 targeting Android, with application state managed through the Provider package and session management handled by Firebase Authentication via auto-refreshed JWTs. The backend source code is maintained in a private repository at `https://github.com/polMarsol/Necodys-App-Backend`, available upon request. Beyond these technical foundations, the application embodies a set of deliberate design decisions oriented towards its two very different users—the clinical specialist and the four-to-five-year-old child—which the remainder of this section documents.

5.5.1 Design Philosophy: a Dual-User Application

NECODYS is conceived as a *decision-support* instrument for the specialist, not as an autonomous diagnostic device, and this principle shapes every interface decision. A single application must serve two users whose needs are almost opposite. The specialist requires a sober, information-dense, professional environment—patient registration, informed-consent capture, an expert-opinion form, and result dashboards. The child requires the converse: a calm, playful, low-text environment that sustains attention and reduces anxiety during a spontaneous-speech task. The application resolves this tension through two visually distinct *modes* within one codebase: clinician-facing screens use a sober blue Material 3 palette and conventional form controls, whereas the child-facing task screen relies on large illustrations, colour, voice guidance, and animation.

A second guiding principle is *ecological elicitation*: rather than constraining the child with structured drills, the tasks invite open, spontaneous discourse around familiar situations (an image, a memory, a short story), maximising the morphosyntactic and narrative signal available to the downstream pipeline. A third principle is *accessibility for pre-literate children*:

instructions are never assumed to be read by the child but are spoken aloud (Section 5.5.4), and the typography is chosen for maximum legibility (Section 5.5.6).

The eight evaluation task types, the gamification approach, and the accompanying pedagogical guidelines were defined by the research group at the Institut de Recerca Biomèdica de Lleida (IRBLleida) that proposed the NECODYS project; the present work implements those specifications and translates them into concrete interaction and visual-design decisions.

5.5.2 Evaluation Flow

Figure 5.1 traces a complete evaluation session. After authenticating (Firebase, verified e-mail), the specialist reaches the *Dashboard*, which lists registered patients alongside an overview chart of their discourse levels. A new evaluation begins with the *New Patient Form*, which records demographic data (identifier, date of birth, gender, habitual language), habitual screen exposure across four device types, free-text observations, and—mandatorily—confirmation of informed consent.

The session then enters the child-facing loop. From the *Task Selection* screen the specialist hands the device to the child for each task; the child completes the *Task Screen* and returns to selection, repeating until all active tasks are done. Per-task progress is persisted locally in `SharedPreferences`, so an interrupted session is not lost and a finished task may optionally be re-recorded. Once every task is complete, the *Evaluation Summary* screen collects the specialist's expert opinion, additional observations, and a manual discourse-level score on a 0–10 slider, then uploads all recordings and triggers the session-level analysis. The analysis runs asynchronously on the backend (Section 5.6); results are reviewed later in the patient-detail view, closing the loop from data capture to clinical report.

5.5.3 Evaluation Task Types

The application administers the eight task types listed in Table 5.2, each designed to elicit a different facet of oral discourse. Task content is not hard-coded: each task is a `TaskDefinition` document fetched from the backend (title, instruction, and optional stimulus images), so tasks can be edited or extended without releasing a new build. The specialist can toggle the visibility of individual tasks from a management screen, and each recording uses a default time budget of 90 seconds. Every task is presented to the child under a short, action-oriented Catalan verb label (the *Child label* column) that names the "mission" in a single familiar word.

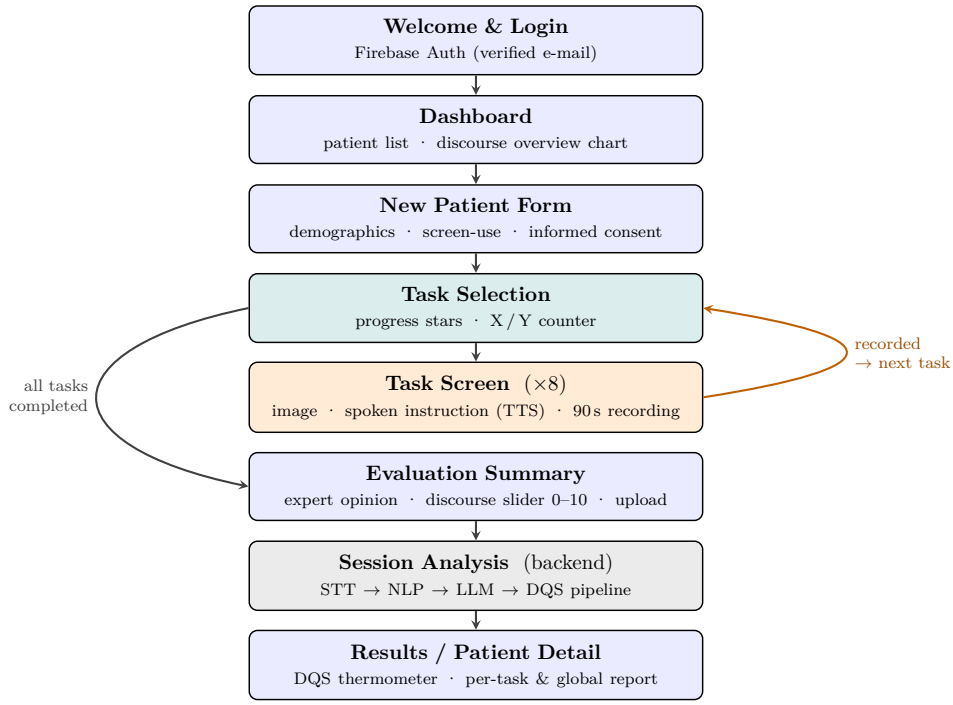


Figure 5.1: Evaluation session flow through the mobile application. Blue boxes are clinician-facing screens, the orange box is the child-facing task screen, and the grey box is the backend analysis. The child-facing loop (orange arrow) repeats the task screen for each of the eight tasks; once all selected tasks are complete, the session advances to the evaluation summary and the asynchronous analysis pipeline.

5.5.4 Child Interaction Behaviours

The task screen is engineered so that a pre-literate child can complete it with minimal adult intervention. On entry, after a short settling delay, the instruction is read aloud by text-to-speech in Catalan (`ca-ES`). Two replay controls are offered: a normal-rate button (*"Escoltar"*) and a slow-rate button (*"Lent"*, marked with a turtle icon) for a child who did not understand the first time; an animated sound-wave indicator replaces the button icon while speech is playing. Text-to-speech stops automatically the moment recording starts, preventing the prompt from contaminating the audio sample. This audio-described delivery benefits both users. For the child, the spoken instruction—together with the normal-rate (*"Escoltar"*) and slow-rate (*"Lent"*) replay—can be re-heard as many times as needed without adult help, so a pre-literate child advances through the task with partial autonomy. For the specialist, it standardises the assessment: the same instruction is voiced identically for every child and every session, which spares

Task type	Child label	Discourse register elicited
`narration`	*"Explica"*	Spontaneous narration of a single image; narrative organisation.
`sequence`	*"Ordena"*	Temporal ordering of an image sequence; connectors and cohesion.
`description`	*"Descriu"*	Description of an object or animal; lexical richness.
`memory`	*"Recorda"*	Autobiographical recall; coherence and self-reference.
`how_to_do`	*"Fes-ho"*	Procedural explanation; functional language and planning verbs.
`story_completion`	*"Continua"*	Open-ended story continuation; cognitive flexibility and connectors.
`qa_game`	*"Respon"*	Rapid question–answer; lexical access and response coherence.
`repetition`	*"Repeteix"*	Sentence repetition with transformation; verbal memory and morphosyntax.

Table 5.2: The eight evaluation task types administered by the mobile application. Each maps to a backend `task_type` and is presented to the child under a single action verb. The taxonomy follows the guidelines of the research group that proposed the project.

the clinician from reading each prompt aloud and removes the variability and potential bias that an adult's own phrasing, intonation or pacing would otherwise introduce into the elicited speech sample.

Recording is captured in M4A format through the `record` package after an explicit microphone-permission check. While recording, a circular countdown ring displays the seconds remaining against the 90-second budget, a concentric ring pulses to signal that the microphone is live, and the child can pause and resume. To sustain engagement, rotating encouragement messages—for example *"Molt bé, segueix!"*—cycle every ten seconds, and on completion a full-screen success overlay with a star and the message *"Ho has fet molt bé!"* confirms the achievement before returning to task selection. Stimulus images are shown either as a single large tappable image (with pinch-to-zoom) or, for sequencing tasks, as a vertically numbered series.

5.5.5 Gamification

Gamification here is a functional requirement rather than decoration: a four-to-five-year-old child must remain motivated and unanxious throughout roughly ninety seconds of unsupported spontaneous speech, repeated across several tasks. The design therefore frames the session as a sequence of short "missions" and provides continuous, legible feedback on progress.

Concretely, the *Task Selection* screen presents a star for each task—outlined when pending, filled gold when complete—together with a gold

progress bar and an "X / Y" counter, giving an at-a-glance sense of advancement. Each task card carries its own coloured icon and verb label, giving every mission a distinct identity; a completed card is visually checked off (tick, strikethrough, and a gold star). When all tasks are done, a celebration banner appears and a prominent trophy-marked button leads to the final report. Within a task, the idle microphone button gently pulses to invite the first tap, and the per-task success overlay rewards completion immediately. All of these elements are implemented directly with Flutter's animation primitives (`AnimationController` and custom overlays); no third-party gamification library is used, keeping the dependency surface small. Critically, the reward schedule is tied to *participation*, not to performance: the child is congratulated for completing a task regardless of its linguistic content, so the gamification never biases the speech sample or penalises a struggling child.

5.5.6 Typography and Visual Identity

Two typefaces explicitly designed around legibility were considered for the application. *OpenDyslexic* is a free typeface whose letterforms carry weighted bottoms and exaggerated, asymmetric shapes intended to reduce the rotation and mutual confusion of similar glyphs for dyslexic readers. *Lexend* is a family designed to improve reading proficiency and reduce visual stress through wide spacing and open, unambiguous forms. Figure 5.2 shows both, each label and sample line set in its own design.

Figure 5.2: Type specimen of the two candidate typefaces, each rendered in its own design: *Lexend*, the family adopted by the application, and *OpenDyslexic*. OpenDyslexic's heavier, asymmetric forms are optimised for dyslexic readers decoding running text, whereas Lexend favours a neutral, evenly spaced rhythm suited to a professional clinical interface.

Lexend was ultimately adopted, applied globally through the `google_fonts` package. The decisive consideration is who actually reads the interface: the child is guided by spoken instructions and is not expected to read the screen (Section 5.5.4), so the on-screen text is read almost exclusively by the specialist. For that reader the priority is a clean, neutral, professional appearance, and Lexend additionally carries an evidence-oriented design rationale around reading proficiency that makes it a thematically coherent choice for a tool situated in the domain of reading difficulty. OpenDyslexic's

distinctive heavy forms, by contrast, are optimised for dyslexic readers decoding long passages—a use case the application does not present. This selection is a reasoned design decision rather than the outcome of a controlled legibility study with the target population.

The visual identity reinforces the dual-user philosophy. A calm primary blue (■ / ■) anchors the clinician-facing screens, while each of the eight evaluation tasks is assigned a distinct accent colour and rounded icon that the child comes to associate with that mission (■■■■■■■■). Throughout, the interface favours Material 3 rounded shapes, generous spacing, and large touch targets appropriate for small hands.

5.5.7 Clinician-Facing Views

Once a session has been analysed, the specialist consults the results without leaving the application. The *Dashboard* summarises the patient cohort and renders an overview chart of discourse levels (`fl_chart`), with each patient's manual level colour-coded. The *Patient Detail* view presents both per-task and global `PipelineResult` documents, including recorded-audio playback, the NLP metrics, the morphosyntactic LLM report, and the discourse quality score surfaced through a dedicated thermometer widget (0–10 with its sub-indicators). To avoid redundant cost, the view requests an existing analysis before triggering a new one, mirroring the idempotent backend behaviour described in Section 6.5.

A note on validation scope is warranted. The task taxonomy, the gamification scheme, and the typographic choices described above are grounded in the guidelines provided by the research group and in established accessibility practice, but they have not yet been empirically validated with real children: no usability study or engagement measurement was conducted within the scope of this dissertation, and the client currently targets Android only. Confirming that these design decisions achieve their intended effect on child engagement and data quality is part of the prospective work outlined in Chapter 7.

5.6 API, Data Storage, and Deployment

Authentication is enforced on every endpoint through a FastAPI dependency that verifies the Firebase JWT and returns HTTP 401 before any database or inference call is made. No credentials are stored server-side.

Data persistence uses MongoDB via the Motor asynchronous driver. Document identifier inputs are validated before query construction. Both per-task and global analysis endpoints check for an existing completed result before invoking any external API, returning the cached document immediately on a cache hit.

The backend is deployed as a single Docker service on port 8000 with automatic restart on host reboot. Two persistent volumes mount audio files and task images independently of image rebuilds. Secrets and model identifiers are injected from a `.env` file via Docker Compose.

Figure 5.3 illustrates the complete evaluation session message sequence, from audio upload through transcription to analysis result delivery.

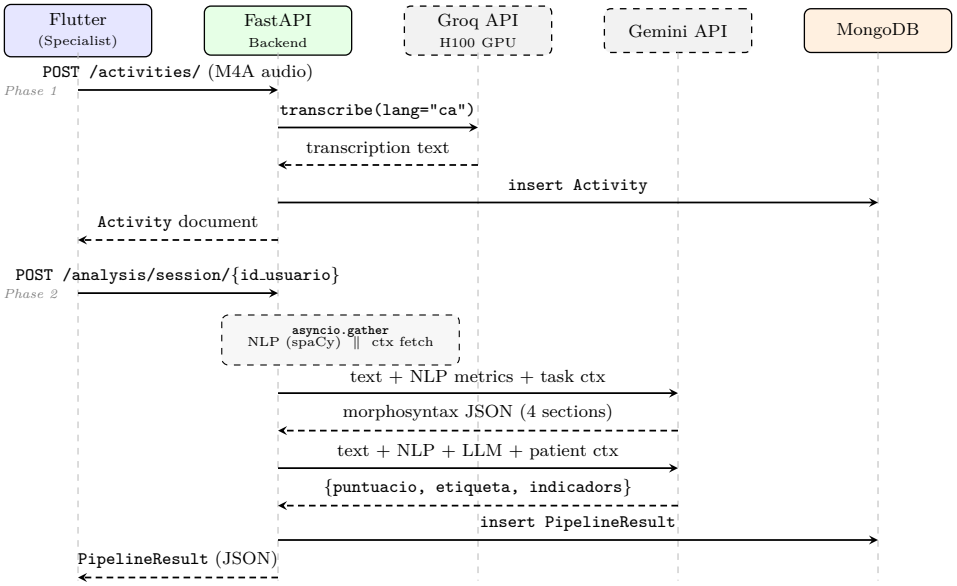


Figure 5.3: Evaluation session message sequence. *Phase 1* (upload and transcription): the mobile client submits an M4A recording; the backend transcribes it via Groq and persists an `Activity` document in MongoDB. *Phase 2* (analysis): NLP feature extraction and context retrieval execute concurrently inside `asyncio.gather` (dashed box); two sequential Gemini calls produce the morphosyntactic report and the discourse quality score; the aggregated `PipelineResult` is persisted and returned to the client. Dashed borders denote external cloud services.

Chapter 6 evaluates the implemented components against the criteria defined in Chapter 3.

Chapter 6

Results

6.1 Speech-to-Text Results

The full STT benchmark is reported in Table 3.1 (Chapter 3). Among the six candidates, `whisper-large-v3` is the only model combining RTF < 1 with WER < 0.20 and DPR(A) $> 60\%$, and was therefore selected as the production model. Its WER advantage over all operationally viable alternatives is statistically significant under Wilcoxon signed-rank tests with Bonferroni correction: versus `whisper-large-v3-turbo` ($p = 0.0005$) and versus `faster-whisper-tiny` and `faster-whisper-base` (both $p < 0.001$). The difference relative to `faster-whisper-medium` was not significant ($p = 0.268$), but that model is operationally excluded by its RTF of 100.5× real-time.

One systematic limitation must be carried forward to the interpretation of subsequent results: Whisper exhibits a normalisation bias that silently removes approximately 33% of consecutive word repetitions present in child utterances (DPR(A) = 67.2%, implying that 32.8% of reference repetitions are erased). Because consecutive-word repetition is the primary disfluency marker in high-risk profiles, the repetition-rate values reported in Section 6.2 are conservative underestimates of the true repetition frequency in Level A speech.

6.2 NLP Feature Discrimination

Table 6.1 reports the mean values of the five primary NLP metrics across the three development corpus quality levels ($n = 40$ children per level). Adult validation sentences ($n = 10$) are excluded from this table to preserve the clinical reference range for the 4–5-year-old population.

All five metrics produce statistically significant monotone gradients across levels. MLU shows the largest effect ($H = 95.9$), with mean utterance length more than doubling from Level A to Level C. Repetition rate is the strongest

Metric	Level A	Level B	Level C	H	p
MLU (words / utterance)	3.23	6.34	9.24	95.9	< 0.001
Type-token ratio (TTR)	0.732	0.815	0.882	42.3	< 0.001
Repetition rate	0.217	0.128	0.054	39.2	< 0.001
Mean tree depth	1.31	2.36	2.88	63.7	< 0.001
Subordination index	0.087	0.200	0.412	12.2	< 0.001

Table 6.1: NLP metrics across development corpus quality levels ($n = 40$ per level, children only). H: Kruskal–Wallis test statistic, which follows a χ^2 distribution with $k - 1 = 2$ degrees of freedom ($k = 3$ quality levels); larger H indicates stronger between-group separation (e.g. $H = 95.9$ for MLU means the three-group difference is far larger than expected by chance, equivalent to a very large effect, whereas $H = 12.2$ for subordination index reflects a smaller but still significant separation). All five metrics are significant discriminators ($p < 0.001$). Repetition rate values are conservative underestimates owing to the Whisper normalisation bias documented in Section 6.1.

absolute discriminator between extreme groups (Level A: 0.217 vs. Level C: 0.054), and its direction is clinically interpretable: high-risk profiles accumulate consecutive repetitions at four times the rate of typical-development profiles. Given the STT normalisation bias, the true separation is likely wider.

Subordination index yields the smallest effect ($H = 12.2$), consistent with the near-chance Cohen $\kappa = 0.014$ for automatic complex-sentence detection reported in Chapter 3: spaCy captures subordination via dependency relations but misses coordination, causing systematic false negatives. Despite this limitation the index remains statistically significant and monotone, reflecting that Level C child speech contains genuinely subordinated clauses that the parser detects.

6.3 LLM Morphosyntactic Analysis Results

Two candidate models were fairly evaluated on the development corpus ($n = 120$ child samples); a third model, `llama-3.3-70b-versatile`, was attempted but excluded due to systematic API rate-limit failures that left 41 of 120 samples unevaluated, making reliable metric computation impossible (see Section 7.5). Table 6.2 reports the two evaluation criteria; cell shading indicates relative performance (green = better, red = worse).

Schema compliance is the primary production criterion: a model that fails to return a valid JSON structure cannot be integrated into the automated pipeline. `gemini-3.1-flash-lite-preview` achieves 100% compliance across all 120 child samples, versus 95.8% for `llama-3.1-8b-instant`.

On clinical discrimination, `gemini-3.1-flash-lite-preview` shows sub-

Model	Schema	Spearman ρ	Latency (s)
`gemini-3.1-flash-lite-preview`	100.0%	0.774	12.1
`llama-3.1-8b-instant`	95.8%	0.386	5.8

Table 6.2: LLM morphosyntactic benchmark ($n = 120$ child samples, two fairly evaluated candidate models). Schema: proportion of responses returning a valid, complete JSON structure. Spearman ρ: rank correlation of predicted lexical richness with proficiency level ($p < 0.001$ for all models). Latency: mean inference time per sample.

stantially stronger performance: Spearman $\rho = 0.774$ versus $\rho = 0.386$ for `llama-3.1-8b-instant` ($p < 0.001$ for both), a difference of 0.388 rank-correlation units. Only the Gemini model exceeds the $\rho \geq 0.5$ threshold considered informative for ordinal stratification on the development corpus. `gemini-3.1-flash-lite-preview` was therefore selected as the production model.

One distributional property of the selected model on this corpus is worth noting: the *high* lexical richness category was never assigned to any of the 120 child samples, yet all ten adult validation sentences were correctly classified as *high*. This is a corpus-level ceiling effect — the development corpus represents 4–5-year-old discourse, which does not exhibit the syntactic complexity required to activate the *high* label — not a defect of the model itself.

6.4 Discourse Quality Score Agreement

The agreement between the automatic DQS and the expert reference score is the primary quantitative result of this work. Spearman rank correlation across all 130 corpus samples was $\rho = 0.904$ ($p < 0.001$). Restricting analysis to the 120 child-only samples yields $\rho = 0.884$ ($p < 0.001$), confirming that the result is not driven by the well-separated adult boundary cases.

Figure 6.1 presents two complementary views of the agreement. The **left panel** is a scatter plot of automatic DQS (y-axis) versus expert reference score (x-axis), colour-coded by quality level. The dashed diagonal represents perfect agreement: points *above* the diagonal mean the automatic system assigned a higher score than the expert (overestimation); points *below* indicate underestimation. The visible cluster structure confirms the ordinal separation: Level A samples concentrate in the lower-left region (low scores on both axes), Level C in the upper-right, and Level B in between, with no distributional overlap between the two extreme groups. The **right panel** is a Bland–Altman plot, where the x-axis is the mean of the two scores $\left(\frac{\text{auto}+\text{expert}}{2}\right)$ and the y-axis is their signed difference (auto − expert): positive values mean the automatic score exceeds the expert's; negative values mean the expert assigned a higher score. The central dashed line marks

the mean bias; the outer dashed lines mark the 95 % limits of agreement (LoA), the interval within which 95 % of future pairwise differences would be expected to fall if the same methods were applied to a new sample. The Bland–Altman analysis therefore quantifies the systematic and random components of disagreement:

- **Mean bias**: +0.113 points, indicating a negligible systematic offset with no clinical significance.
- **Standard deviation of differences**: 1.174 points.
- **95% limits of agreement (LoA)**: $[-2.20,\ +2.42]$.
- **Proportion within 1 point**: 65.4%; within 2 points: 90.8%.

Table 6.3 summarises the per-level means to contextualise the LoA.

Level	**Auto DQS (mean ± SD)**	**Expert (mean ± SD)**
A (high risk, $n = 40$)	2.79 ± 1.01	2.59 ± 0.80
B (intermediate, $n = 40$)	6.16 ± 0.60	5.49 ± 1.13
C (typical, $n = 40$)	7.28 ± 0.69	7.88 ± 1.12

Table 6.3: Automatic DQS and expert reference score by quality level ($n = 40$ children per level). The two extreme groups are clearly separated by approximately 4.5 scale units, with no distributional overlap between them.

The LoA span of 4.6 points on a 0–10 scale confirms that the automatic score and the expert reference score track the same ordinal signal on the development corpus, though not as precision measurements. The discriminative range documented here—a Level A mean DQS of 2.79 versus a Level C mean of 7.28, with no distributional overlap between the extremes—provides a concrete quantitative basis for the prospective clinical validation study required before any deployment in referral contexts.

The near-zero bias (+0.113) confirms that the scoring scale is well-calibrated against the expert reference. The independence constraint described in Chapter 3 — whereby the expert reference score is never exposed to any LLM prompt — ensures that this agreement reflects genuine concordance rather than circular self-reporting.

6.5 Pipeline Technical Performance

Table 6.4 reports the measured latency per pipeline stage on the development server under normal network conditions.

The STT and NLP stages execute concurrently; the bottleneck is the cloud STT call, whose duration scales with audio length at RTF = 0.47 (i.e.

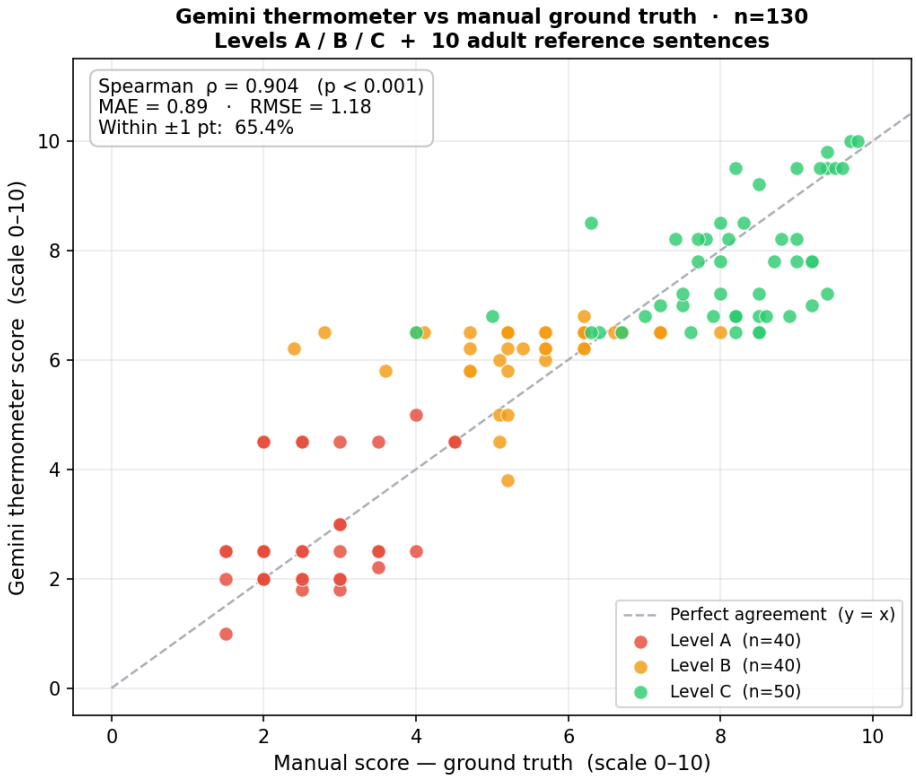


Figure 6.1: **Left:** scatter plot of automatic DQS versus expert reference scores ($n = 130$), colour-coded by quality level; dashed diagonal = perfect agreement (points above diagonal = automatic overestimates; points below = underestimates). **Right:** Bland–Altman plot (x = mean of both scores; y = auto − expert; central line = mean bias; outer lines = 95 % LoA). The ten adult validation sentences are shown separately (Level ADULT). Spearman $\rho = 0.904$ ($p < 0.001$).

faster than real time). The two sequential Gemini calls each add approximately 12 s, yielding the 25–35 s total for a typical 30 s evaluation clip.

Two robustness properties are worth reporting as measurable system characteristics. *Idempotent caching*: if a `PipelineResult` document with `estat = completed` already exists for a given activity, the pipeline returns the cached result immediately (<1 s) without issuing any LLM calls. Re-execution is permitted when the stored state is `error` or `partial`, enabling correction without data loss. *Graceful degradation*: each pipeline module returns `None` on failure rather than propagating exceptions; a partial result is stored with `estat = partial` and the clinical report marks the unavailable sections explicitly. Both properties were verified during integration testing on the development server.

Stage	**Execution**	**Measured latency**
STT — Groq `whisper-large-v3`	parallel	RTF 0.47× (≈14 s per 30 s clip)
NLP — spaCy `ca_core_news_sm`	parallel	< 1 s (CPU-bound)
LLM morphosyntax — Gemini	sequential	≈12 s mean
LLM thermometer — Gemini	sequential	≈12 s mean
Total (end-to-end, measured)		**25–35 s**

Table 6.4: Per-stage pipeline latency under normal network conditions. STT and NLP run concurrently via `asyncio.gather`; the two LLM calls run sequentially because the thermometer receives the morphosyntactic output as input.

These operational characteristics confirm that the system architecture satisfies the technical feasibility criterion established in Objective 4 (Section 1.4).

Chapter 7

Discussion

7.1 Scope of This Work

The original objective of this dissertation was to develop a prototype support system for the early identification of oral language risk indicators associated with developmental dyslexia in pre-literate children aged four to five years. The present work delivers the first two phases of that goal: system construction—a fully functional end-to-end pipeline comprising mobile data collection, cloud speech recognition, NLP feature extraction, and LLM-based morphosyntactic profiling—and internal calibration against a manually authored development corpus. Clinical validation, the phase that would establish whether the system's discriminative properties transfer to real child speech, lies outside the scope of this dissertation and constitutes the necessary next step.

The results presented in Chapter 6 are promising but not conclusive. Promising, because all five primary NLP metrics produce statistically significant monotone gradients across the three quality levels ($p < 0.001$, Kruskal–Wallis), the discourse quality score (DQS) reaches Spearman $\rho = 0.904$ against expert reference scores, and end-to-end processing completes in under 35 s—confirming that the proposed architecture is technically viable. Not conclusive, because all evidence derives from a manually authored corpus: no real child recordings were collected, quality level assignments were made by a single researcher, and no comparison against any clinical reference standard has been performed. These constraints are inherent to the scope of a bachelor's dissertation and are addressed as limitations in Section 7.5.

7.2 Comparison with the Literature

The discriminative pattern observed across the three quality levels is consistent with the broader body of longitudinal research on pre-literacy oral language. Scarborough [3] documented that children later diagnosed with

dyslexia already exhibit shorter mean length of utterance and lower lexical diversity at age 2.5 years, well before formal literacy instruction begins. The MLU gradient observed here—3.23, 6.34, and 9.24 words per utterance for levels A, B, and C respectively—mirrors this finding at the discourse level: the system correctly assigns its lowest scores to profiles characterised by telegraphic, fragmented utterances, and its highest scores to profiles showing embedded subordination and varied vocabulary.

Repetition rate emerges as the strongest absolute discriminator between extreme groups (Level A: 0.217 vs. Level C: 0.054; $H = 39.2$, $p < 0.001$). This finding is consistent with Catts et al. [4], who showed that oral language disruptions—including disfluencies and restricted syntactic structures—are reliable early predictors of reading difficulties, and with Bishop and Snowling [14], who argue that language deficits in children at risk for dyslexia extend beyond phonological processing to broader morphosyntactic and discourse organisation. The clinical relevance of this marker is strengthened by the fact that it is entirely text-based and therefore independent of the acoustic analysis module currently implemented as a Phase 2 stub.

The DQS agreement ($\rho = 0.904$, mean bias $+0.113$, LoA $[-2.20, +2.42]$) is comparable in magnitude to inter-rater reliability figures reported for structured clinical instruments in this developmental domain [1]. However, a direct numerical comparison is not warranted: the expert reference scores were provided by the researcher rather than by independent credentialled clinicians, and the corpus does not consist of real patient recordings.

One distributional property of the LLM component is worth contextualising: `gemini-3.1-flash-lite-preview` never assigns the *high* lexical richness category to any of the 120 child samples, while correctly classifying all ten adult validation sentences as *high*. This is not a model deficiency but a corpus-level ceiling effect. The discourse of neurotypical four-to-five-year-old children is inherently limited in morphosyntactic complexity relative to adult speech [34]; the model is responding correctly to the input it receives. In a clinical deployment, this property implies that the *high* category may only be reached by the most advanced children or older age groups—a hypothesis that prospective data would test.

A point of divergence from existing approaches concerns both assessment modality and target population. Among the systems reviewed in Chapter 2, only EarlyBird [18] (ages 4–7) and Berni [19] (ages 4–6) are validated for pre-reading children, and both rely on structured, task-constrained formats—phoneme-awareness items, non-word repetition, and rapid automatised naming, rather than on spontaneous discourse. The remaining tools (Dytective [20], Lexplore [21], SICOLE-R [22]) presuppose baseline reading ability and are therefore inaccessible to the preschool population, while Catalan-context efforts such as the PREVENIR oral-language programme [35] target intervention rather than screening. None of these systems analyses spontaneous oral discourse in Catalan. NECODYS, by contrast, elicits spontaneous

oral discourse across eight ecologically grounded task types. This modality offers higher ecological validity and greater sensitivity to the full range of morphosyntactic and narrative-organisation deficits, but introduces additional variability that standardised task-based instruments explicitly suppress. The trade-off between ecological validity and measurement standardisation is a core design tension that prospective clinical studies would need to quantify.

7.3 Positioning within Computational and Technological Approaches

Methodologically, NECODYS aligns with an emerging body of work that applies natural language processing and machine learning to the early identification of language disorders. Lammert et al. [36] survey precisely this paradigm and identify its central challenges—small annotated corpora, the scarcity of child speech data, and the demand for interpretable outputs—all of which the present work encounters directly. Using a large language model for structured morphosyntactic profiling is consistent with recent demonstrations that LLMs can be fine-tuned and systematically evaluated for speech-pathology tasks [37] and with broader assessments of artificial intelligence in neurodevelopmental screening [27]; recent multimodal detection frameworks [38] pursue the same early-detection goal through complementary signal sources. NECODYS contributes a concrete end-to-end instantiation of this computational paradigm for Catalan preschool discourse, a language and demographic combination not previously addressed.

The speech-recognition layer must be read against the documented performance gap between adult and child automatic speech recognition. Whisper-family models are trained predominantly on adult speech, and recent benchmarks report markedly higher error rates on children's speech together with targeted remedies such as Kid-Whisper [39, 40]. This literature contextualises the normalisation bias measured in Section 6.1: the silent deletion of roughly one third of consecutive repetitions in Level A speech is a particular manifestation of a known, systematic adult-centric bias rather than an idiosyncrasy of this deployment, and it reinforces the case for child- and Catalan-tuned recognisers in future iterations.

Finally, the bilingual reality of the target population warrants explicit discussion. Many children in Catalonia grow up with simultaneous exposure to Catalan and Spanish, and phonological-processing profiles in bilingual children with and without dyslexia diverge from monolingual norms [41]. Cross-language interference—illustrated by the castilianised constructions in the corpus's bilingual-interference profile—poses a dual challenge: it complicates automatic transcription, for which Catalan–Spanish code-switching ASR remains an open research problem [42], and it confounds the linguis-

tic interpretation of reduced morphosyntactic complexity, which may reflect second-language acquisition rather than a genuine risk marker. Disentangling bilingual interference from true risk indicators is an essential requirement for any clinical deployment in this sociolinguistic context.

7.4 NECODYS as a Research Platform

The findings position NECODYS as a technological platform prepared to study and prospectively validate candidate linguistic risk markers, rather than as a validated clinical screening tool. The system provides the technical infrastructure—mobile data collection, calibrated DQS, structured result storage—required to conduct the prospective study that would answer the open clinical question: do the discriminative properties observed on the development corpus transfer to real child speech, and at what sensitivity and specificity?

7.5 Limitations

- **Manually authored development corpus.** All 130 speech samples were authored by the researcher to represent three discourse quality levels; no real child recordings were collected. Results constitute internal calibration evidence for the system's discriminative range, not validated clinical performance. Generalisation to authentic child speech requires a prospective study with appropriate ethics committee approval. Future work should build on real child-speech corpora such as MyST [43], one of the largest open corpora of children's conversational speech, to replace manually authored material with authentic developmental data.
- **Single annotator.** Corpus quality level assignments and expert reference scores were both provided by the researcher. Independent inter-rater validation with credentialled speech-language pathologists is required before the DQS scale can serve as an objective clinical reference.
- **Acoustic module absent.** Phase 2 features—voice activity detection, speech rate, and pause analysis via `librosa`—are not yet implemented; all `AudioMetrics` fields are currently `None`. Present results reflect text-based analysis only. Acoustic disfluency markers are expected to further strengthen discrimination, particularly at Level A, where the Whisper normalisation bias already suppresses repetition-rate values.
- **Whisper normalisation bias.** Approximately 33 % of consecutive word repetitions in Level A speech are silently removed during tran-

scription (DPR(A) = 67.2%), making repetition-rate values conservative underestimates of true disfluency frequency in real child speech. This bias propagates to all downstream metrics that depend on transcription fidelity.

- **Restricted STT candidate space.** Catalan-specific ASR models (`projecte-aina/whisper-large-v3-ca-3catparla`, `BSC-LT/whisper-bsc-large-v3-cat`, `softcatala/wav2vec2-large-xlsr-catala`) could not be evaluated for two compounded reasons: the HuggingFace Serverless Inference API was not activated for any of these models at the time of evaluation, and running them locally requires dedicated GPU hardware that was not available in this project. The STT comparison is therefore restricted to the Whisper architecture; a Catalan-native model may yield higher DPR and lower WER on child speech, but this hypothesis remains untested. Catalan has only recently reached high-resource status in open speech corpora [44], and dedicated work on Catalan–Spanish code-switching ASR [42] indicates that language- and code-switching-aware recognisers could materially reduce transcription error on bilingual child speech.
- **Incomplete LLM comparative study.** Three of six planned LLM candidates could not be evaluated owing to free-tier API quota exhaustion; the selected model may not represent the globally optimal choice across the full candidate space.
- **Absence of research funding.** All API-based experiments were conducted using free-tier allocations, as no external funding was available for this dissertation. This constraint had direct consequences for experimental completeness: three of six planned LLM candidates could not be evaluated due to daily Gemini quota exhaustion, and Groq inference API rate-limit failures introduced stochastic errors in the `llama-3.3-70b-versatile` benchmark (41 of 120 samples returned failures). A funded study with paid API credits would enable a complete and reproducible comparative evaluation.
- **Regulatory approvals for child data collection not obtained.** Conducting evaluation sessions with real children in school settings requires completing a sequential regulatory process in Catalonia: institutional endorsement from the Institut de Recerca Biomèdica de Lleida (IRBLleida), followed by ethics committee clearance, and finally authorisation from the Departament d'Educació de la Generalitat de Catalunya for school-based data collection. The timeline of a bachelor's dissertation did not allow completion of this regulatory chain. This is the primary reason the corpus consists entirely of manually authored samples; obtaining all three approvals is an explicit prerequisite for any future clinical data collection phase.

Chapter 8

Conclusions & Future Work

8.1 Achieved Objectives

O1 — Functional prototype. *Fully met.* A fully functional end-to-end prototype was developed and deployed. The Flutter 3 mobile application provides 12 screens and guides the specialist through eight structured evaluation task types—image narration, sequence explanation, object/animal description, procedural explanation (*how would you do it?*), autobiographical memory recall, story completion, rapid question–answer, and phrase repetition with transformation—covering the full range of oral discourse registers relevant to pre-literacy assessment. Catalan text-to-speech (`flutter_tts`, locale `ca-ES`) automatically reads task instructions to the child at session start, with a slow-rate replay option for children who need repetition. The FastAPI backend exposes four REST endpoints (task-level and session-level analysis and retrieval), with Firebase JWT authentication, asynchronous MongoDB storage via Motor, and Docker containerised deployment with automatic restart. For each recorded audio clip, the pipeline returns a structured `PipelineResult` document containing the Whisper transcription, over 20 NLP metrics, a four-section morphosyntactic LLM report, and a 0–10 discourse quality score within a single request–response cycle.

O2 — Metric discrimination. *Fully met on the development corpus.* All five primary NLP metrics produce statistically significant monotone gradients across the three quality levels (Kruskal–Wallis, all $p < 0.001$). MLU yields the largest effect ($H = 95.9$): mean utterance length nearly triples from Level A (3.23 words) to Level C (9.24 words), reflecting the contrast between telegraphic utterances such as *"Nena nena corre. Gosset gosset menja."* and structurally complete sentences such as *"Hi ha una nena que corre al parc i té un gos gran que li llança una pilota."*. Repetition rate is the strongest absolute discriminator between extreme groups (Level A: 0.217 vs. Level C: 0.054, $H = 39.2$), consistent with longitudinal evidence that disfluency markers are among the earliest reliable

predictors of reading difficulties in pre-literate children [4]. At the LLM stage, `gemini-3.1-flash-lite-preview` achieves Spearman $\rho = 0.774$ ($p < 0.001$) for lexical richness discrimination—the only model among the three evaluated to exceed both the $\rho \geq 0.5$ informativeness threshold and 100 % schema compliance. The qualification *on the development corpus* is deliberate: all samples are manually authored and the generalisation of these gradients to real child speech has not been tested.

O3 — Calibrated discourse quality score. *Met within the scope of the development corpus.* The DQS was designed with an explicit independence guarantee: the expert reference score is never exposed to any LLM prompt, ensuring that the measured agreement reflects genuine concordance rather than circular self-reporting. Spearman rank correlation on the full corpus ($n = 130$) reached $\rho = 0.904$ ($p < 0.001$); restricting the analysis to the 120 child-only samples yields $\rho = 0.884$ ($p < 0.001$), confirming that the result is not inflated by the well-separated adult boundary cases. Bland–Altman analysis reveals a negligible systematic bias of +0.113 points and 95 % limits of agreement of [−2.20, +2.42], with 90.8 % of samples falling within 2 points of the expert reference. In concrete terms, the system assigns a mean DQS of 2.79 to Level A profiles (telegraphic, high-repetition utterances) and 7.28 to Level C profiles (structured, subordinated discourse), with no distributional overlap between these extremes. This calibration is preliminary: the expert reference scores were provided by a single researcher, and cut-off thresholds for clinical referral require validation against independent raters and real patient populations.

O4 — Technical feasibility. *Fully met.* Full pipeline performance data are reported in Section 6.5. End-to-end processing completes in 25–35 s for a typical 30 s evaluation clip under normal network conditions: the STT and NLP stages execute concurrently (Groq `whisper-large-v3` at RTF = 0.47; spaCy in under 1 s), followed by two sequential Gemini calls of approximately 12 s each. Idempotent result caching ensures that re-querying a completed analysis costs less than 1 s without issuing any LLM calls, and per-module graceful degradation allows the pipeline to store a partial result when an individual component fails rather than discarding the entire session. The acoustic analysis module is architecturally integrated as a Phase 2 stub with a fully defined interface; replacing the stub body with a `librosa` implementation requires no structural changes to the pipeline or data model.

O5 — Integrated evaluation workflow. *Fully met.* Beyond the mere existence of a functional prototype (O1), the application was designed as a single coherent workflow that unifies the entire evaluation behind one specialist-facing tool, described in detail in Section 5.5. The same application serves two distinct experiences: a sober, form-driven interface for the specialist and a playful, gamified one for the child. The specialist registers a patient—demographics, screen-exposure habits and informed consent—after which the app guides the child through the structured session: each of the

eight task types is introduced automatically by Catalan text-to-speech (with a slow-rate replay), recorded with on-screen encouragement and a success animation, and tracked by a star-based progress indicator that frames the tasks as a sequence of completable "missions". On completion the specialist records an expert opinion and a manual discourse-level rating, and the session audio is uploaded and analysed automatically; results are returned to the patient-detail view as a discourse-quality thermometer, per-task and session-level linguistic reports, and a history of previous sessions. Analysis triggering is idempotent—a completed session is served from cache rather than re-invoking the language models—which closes the loop from data capture to clinical result within a single tool. This objective concerns the integration and usability of the workflow rather than its clinical validity: the design has not yet been evaluated with real specialists or children in a field setting, so its usability claims rest on design rationale rather than on a user study.

8.2 Future Work

Six development phases are identified to advance NECODYS toward clinical deployment.

Phase 2 — Acoustic feature extraction. The `audio_processor.py` stub will be replaced with a `librosa`-based implementation extracting speech rate, pause count, mean pause duration, and voiced/unvoiced ratio. The `AudioMetrics` dataclass and module interface are already defined; no pipeline restructuring is required.

Phase 3 — Institutional approvals and research funding. Two prerequisites must be completed before any real-patient data collection can begin. First, the regulatory approval chain must be followed: institutional endorsement from the Institut de Recerca Biomèdica de Lleida (IRBLleida), ethics committee clearance, and authorisation from the Departament d'Educació de la Generalitat de Catalunya for school-based data collection. Second, a research funding application is required to cover paid API access—enabling the LLM candidates that could not be fairly evaluated (three Gemini models excluded due to free-tier quota exhaustion, and `llama-3.3-70b-versatile` excluded due to rate-limit failures) to be properly assessed—and to support the infrastructure costs of a prospective clinical study.

Phase 4 — Clinical validation. Once Phase 3 prerequisites are satisfied, a prospective study with real child recordings from speech-language pathology clinics, independent expert raters with established credentials in speech-language pathology, and comparison against standardised instruments (CTOPP-2—Comprehensive Test of Phonological Processing, second edition; DYSC) will establish sensitivity, specificity, and clinical cut-off thresholds for the DQS scale.

Phase 5 — Unsupervised pattern discovery. Applying clustering algorithms (K-means, HDBSCAN) and dimensionality reduction (t-SNE, UMAP) to the extracted feature vectors may reveal natural patient groupings beyond the three-level schema. Autoencoder-based representations offer a complementary route to discovering latent discourse phenotypes.

Phase 6 — Integrated risk indicator. The current thermometer already incorporates patient age, habitual language, and specialist observations as contextual inputs. The next step is to extend the scoring model with additional patient-level risk factors — screen exposure and parental dyslexia history — and to calibrate the composite score against the standardised clinical diagnoses introduced in Phase 4, completing the transition from a screening support tool to a clinical decision-support system.

The results collectively establish NECODYS as a technically feasible and scientifically grounded technological platform prepared to study and validate future linguistic risk markers—the seed of a clinical tool that requires prospective validation with real child populations. The discriminative properties documented on the development corpus provide a concrete quantitative basis for the clinical validation study that will determine whether these properties transfer to real child speech.

Appendix A

Application Screenshots

This appendix presents the implemented NECODYS mobile application as a step-by-step walkthrough of a complete evaluation session. The screenshots follow the exact order in which a specialist and child move through the application, from first launch to the review of the analysis results, and include representative behavioural, error and result states. Discrete dividers mark the phases of the session; the step numbering runs continuously throughout.

Access and clinician home

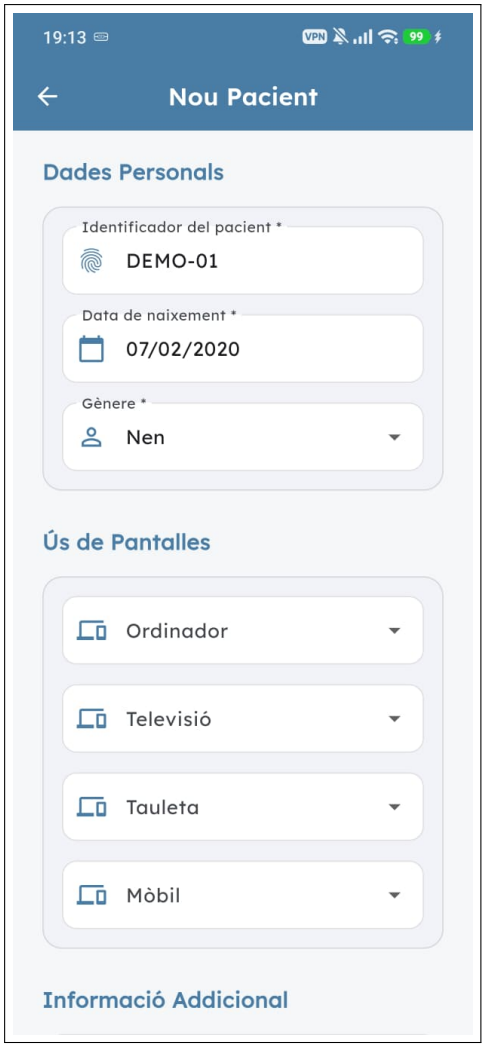
1. The application opens on the welcome screen; the specialist taps *"Començar"* to begin.

2. The specialist signs in with e-mail and password (Firebase authentication).

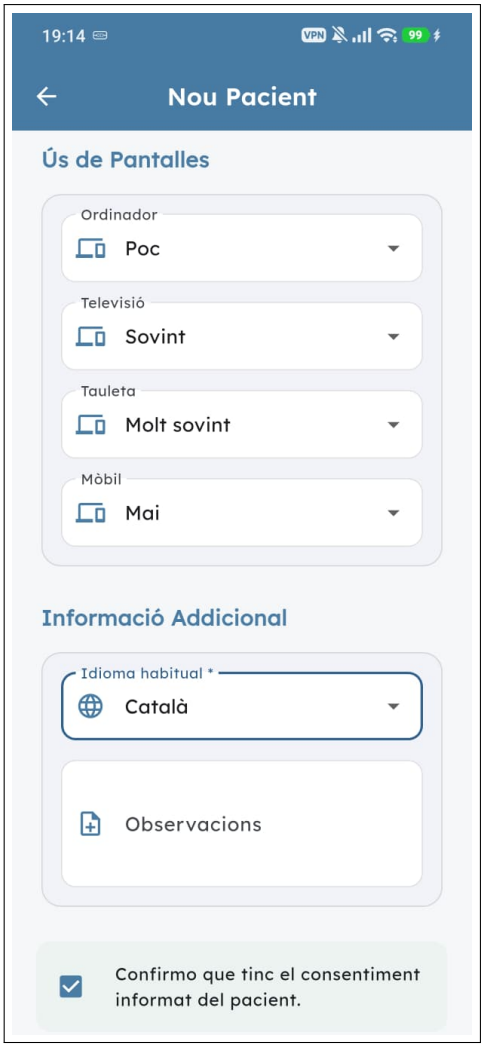
3. Invalid credentials are rejected with an error message.

4. The dashboard shows the discourse-level evolution chart and the list of registered patients.

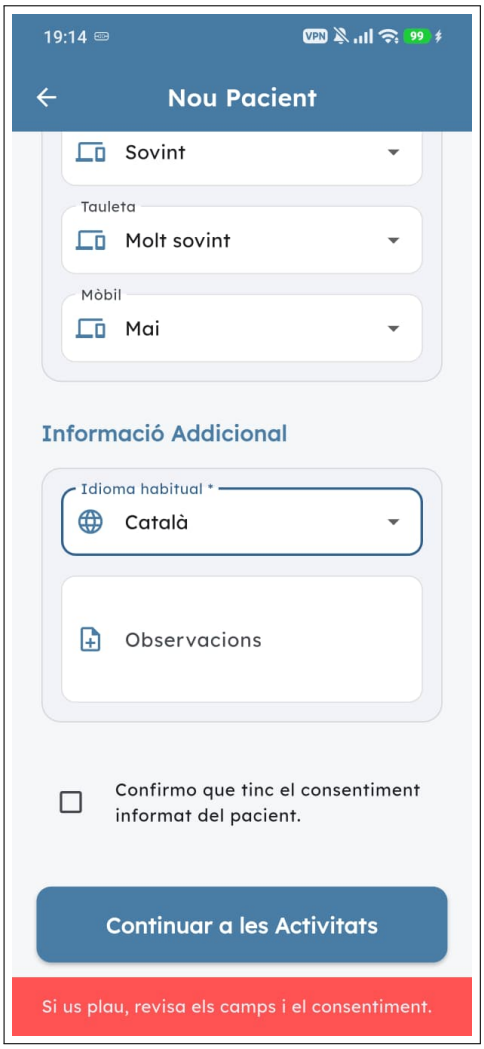
Patient registration



5. A new patient is registered: identifier, date of birth and gender.



6. Screen-use habits and language are recorded and informed consent is confirmed.



7. Submitting without consent or a required field triggers a validation message.

Task selection and gamification

8. The evaluation begins with all eight tasks pending (0 / 8).

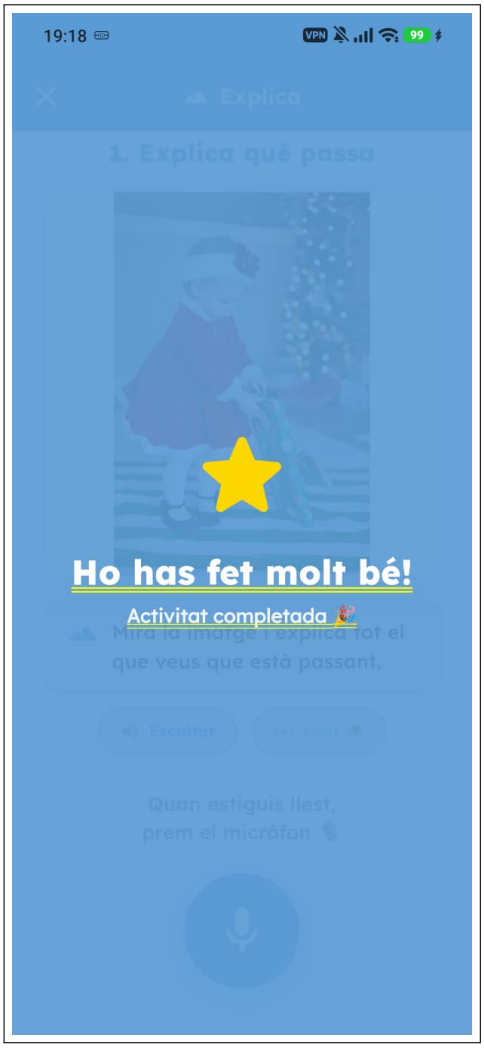
9. When every task is done, all stars are filled and the final-report button appears.

10. The specialist can enable or disable individual tasks.

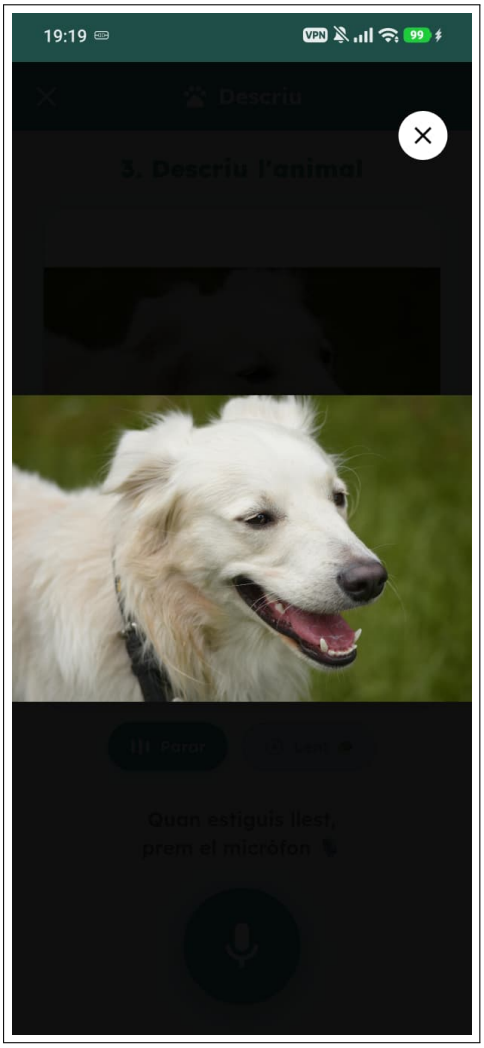
Child task screen and interaction

11. A task opens with its image and spoken-instruction controls (*"Escoltar"* / *"Lent"*); the child taps the microphone when ready.

12. While recording, a countdown ring, an encouragement message and pause/stop controls are shown.



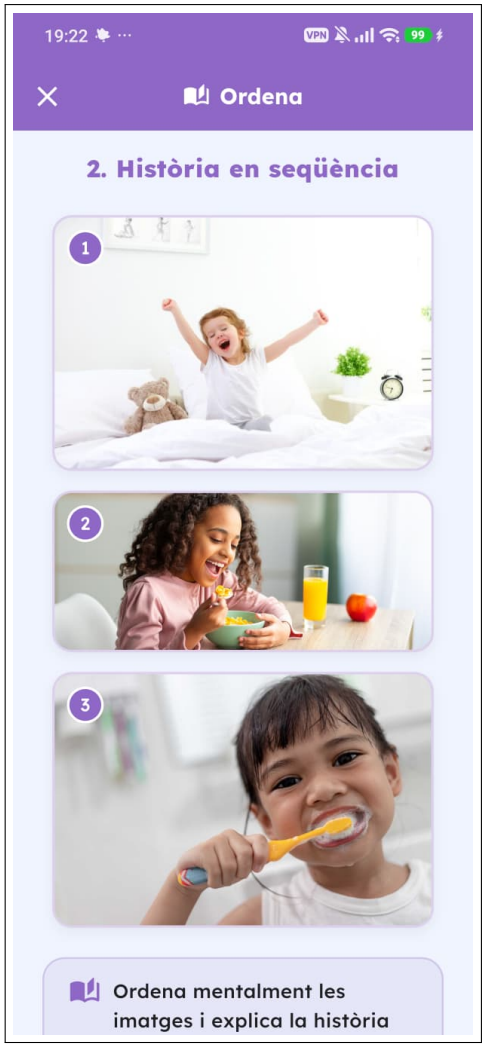
13. On completion, a success overlay rewards the child.



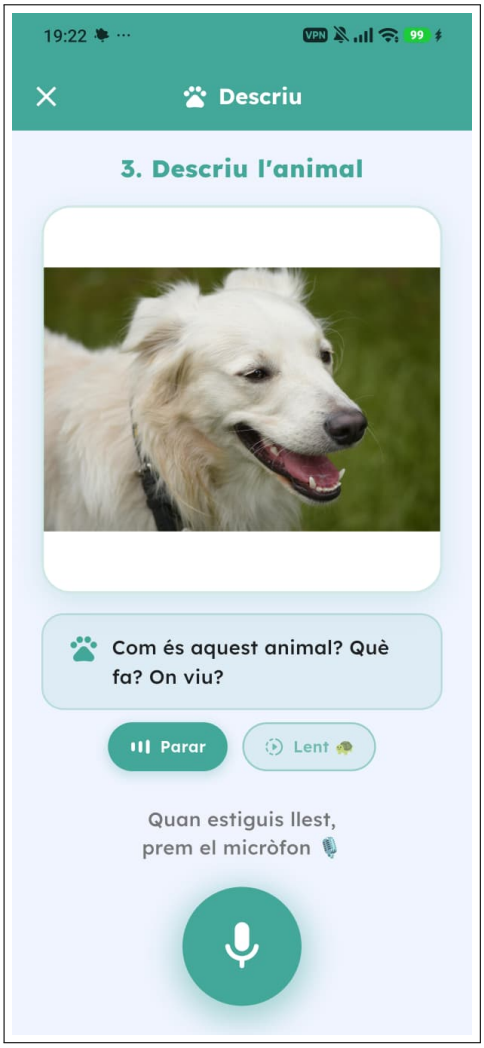
14. The stimulus image can be opened full-screen.

THE EIGHT TASK TYPES

15. Narration task — *"Explica"*.

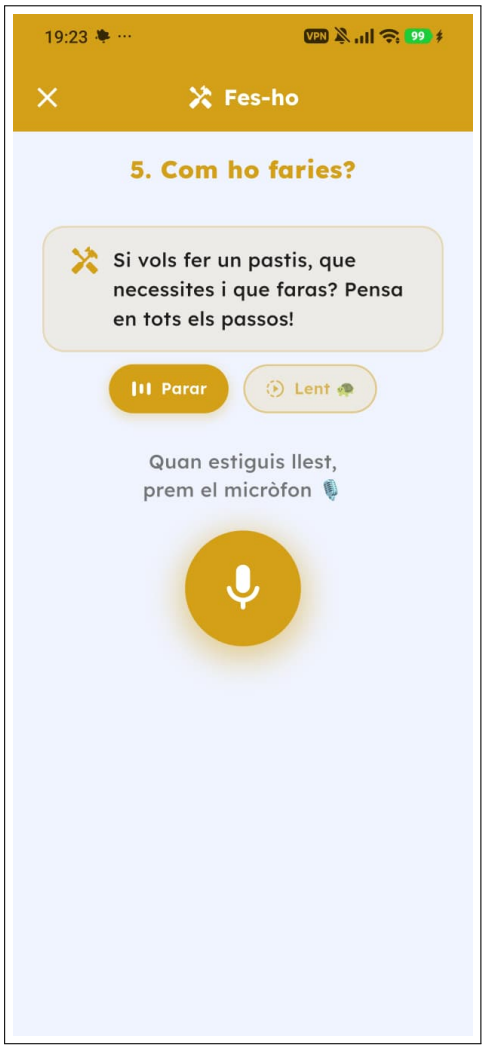


16. Sequencing task — *"Ordena"* — with a numbered image series.

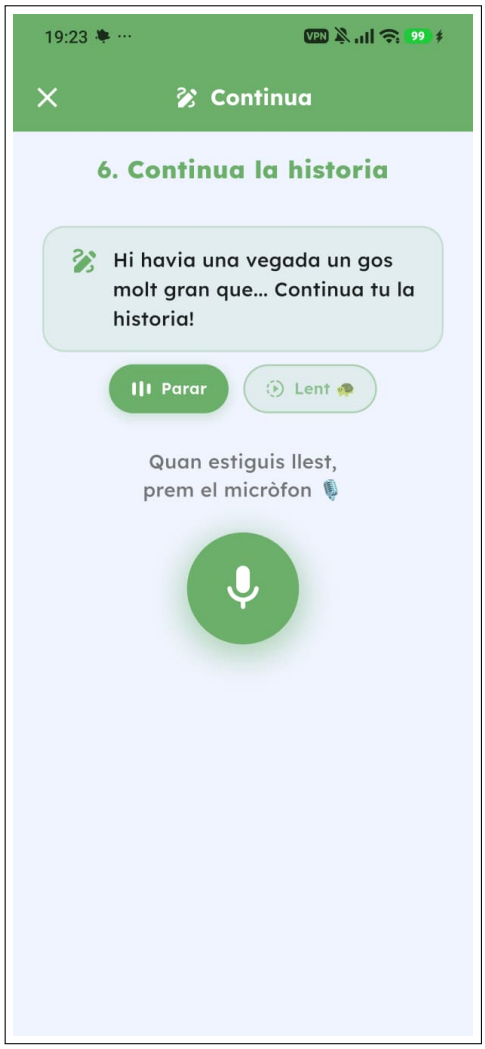


17. Description task — *"Descriu"*.

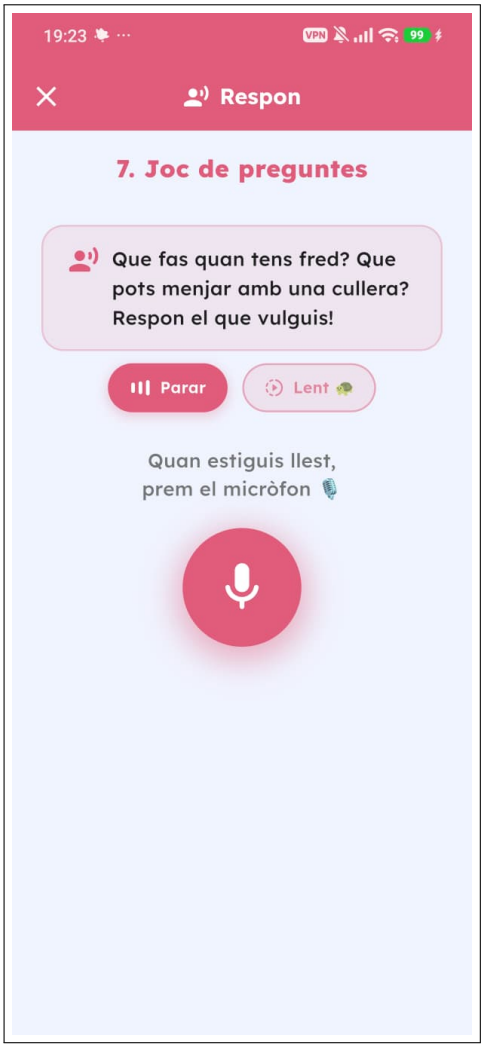
18. Autobiographical recall task — *"Recorda"*.



19. Procedural task — *"Fes-ho"*.



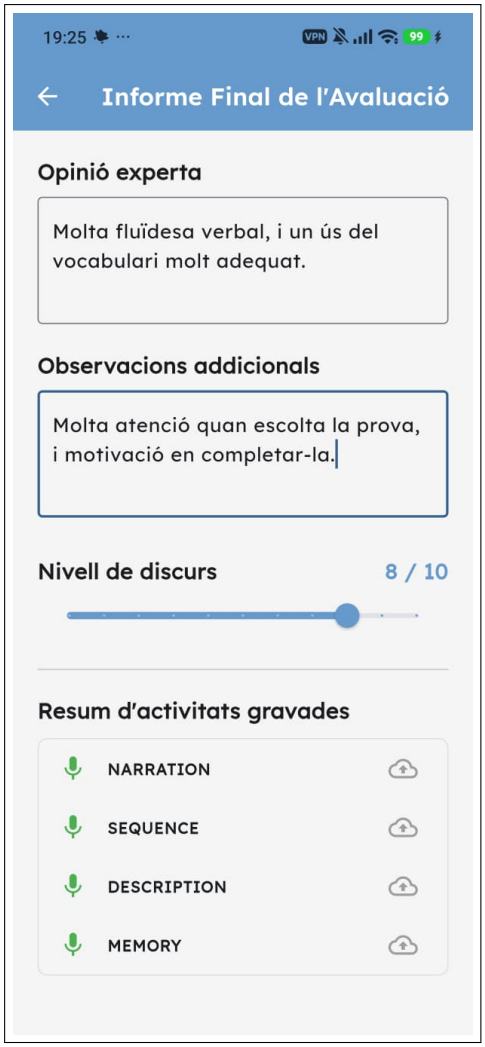
20. Story-completion task — *"Continua"*.



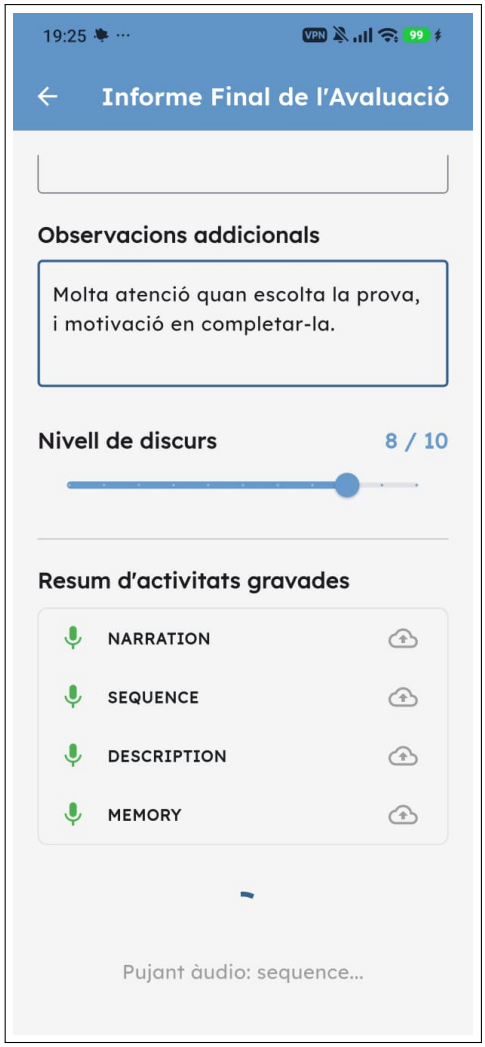
21. Rapid question–answer task — *“Respon”*.

22. Sentence-repetition task — *“Repeteix”*.

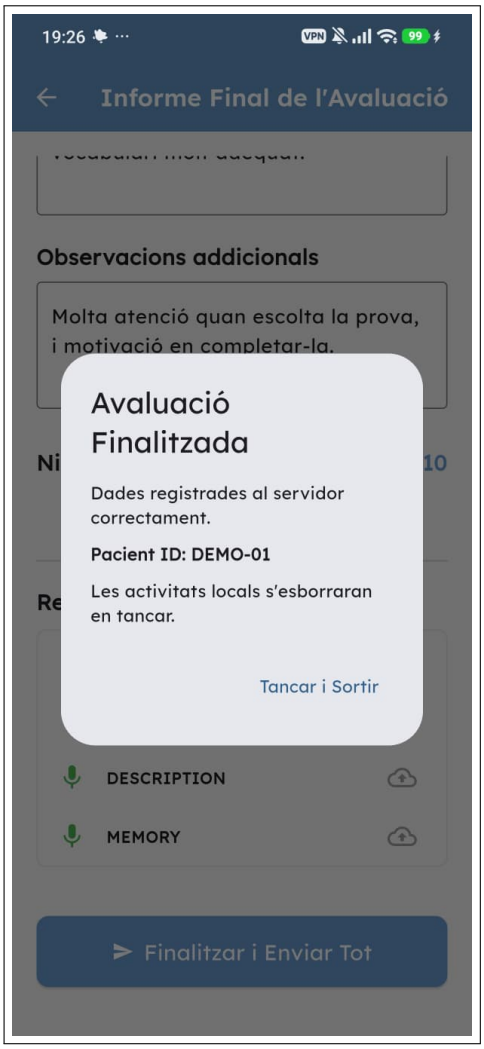
Session summary and submission



23. The session summary collects the expert opinion, observations and the 0–10 discourse-level score.



24. On submission, each recording is uploaded and the session analysis is triggered.

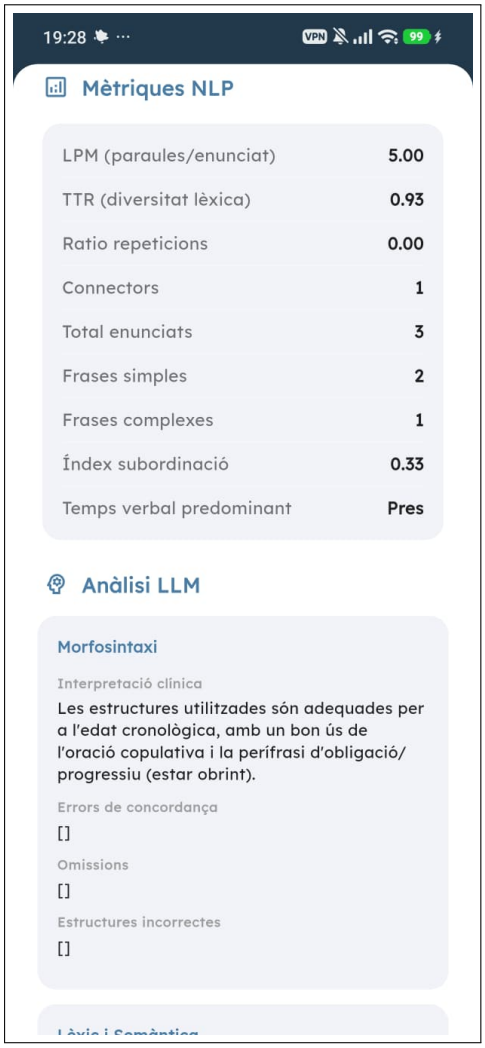


25. A confirmation dialog closes the session.

26. Patient detail (*Current* tab): clinical status, screen-use habits and recorded activities.

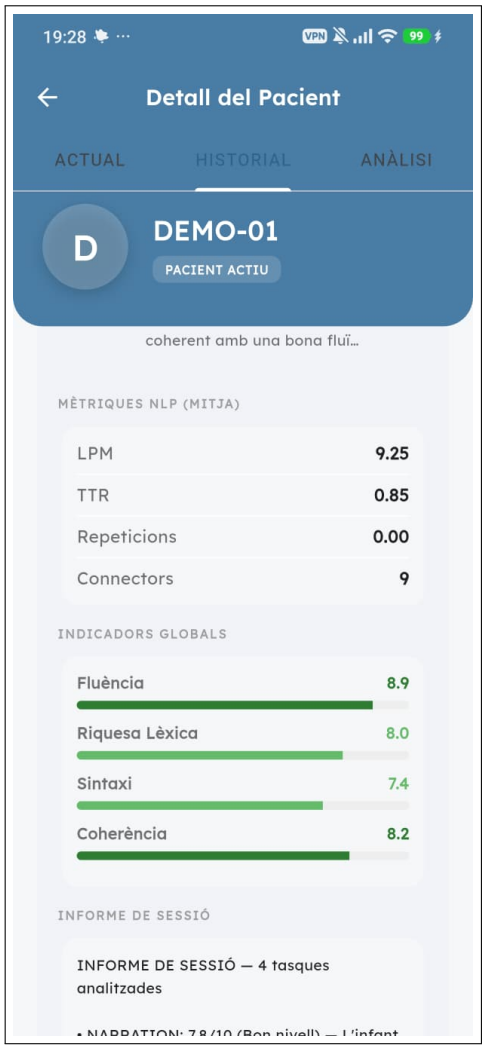
27. Per-task transcriptions with audio playback.

28. Per-task discourse quality score on a 0–10 thermometer with its four sub-indicators.

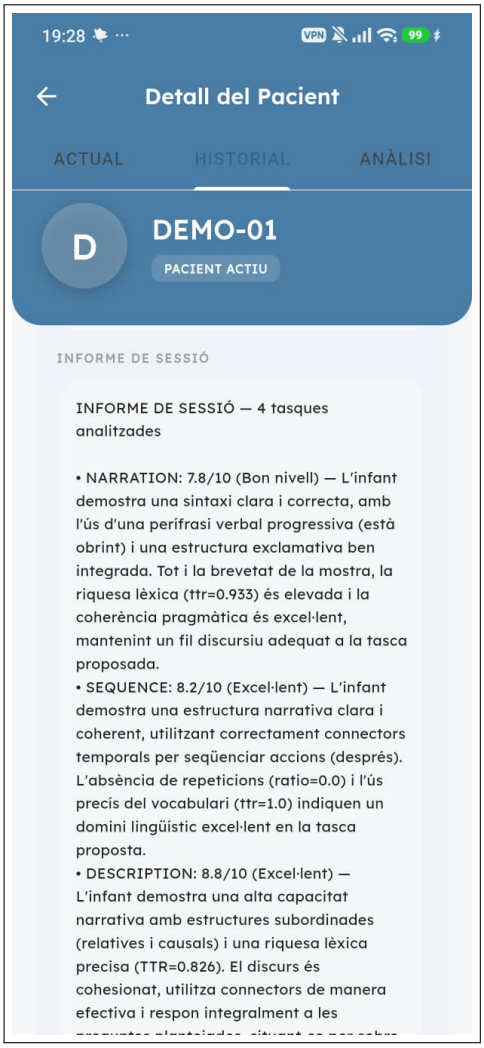


29. Per-task NLP metrics and the morphosyntactic LLM analysis.

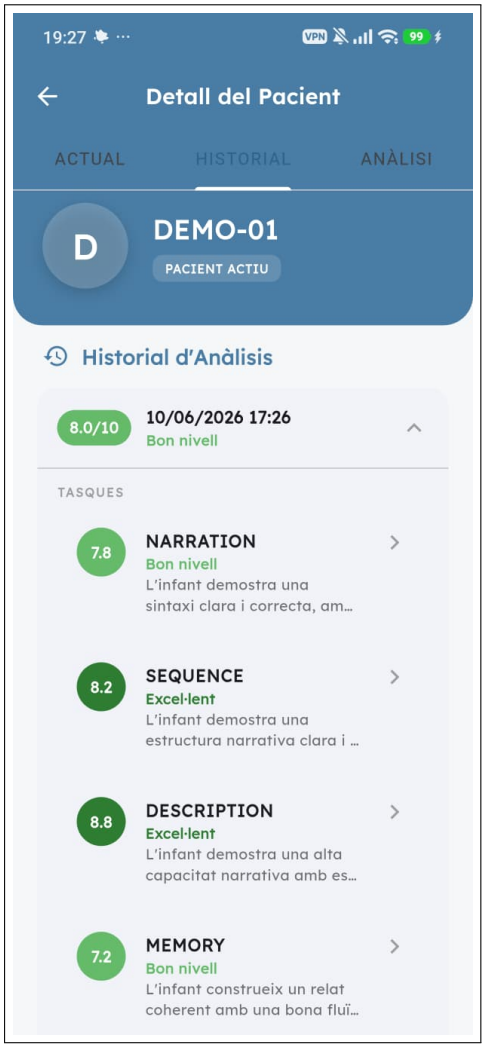
30. LLM lexical–semantic and pragmatic analysis sections.



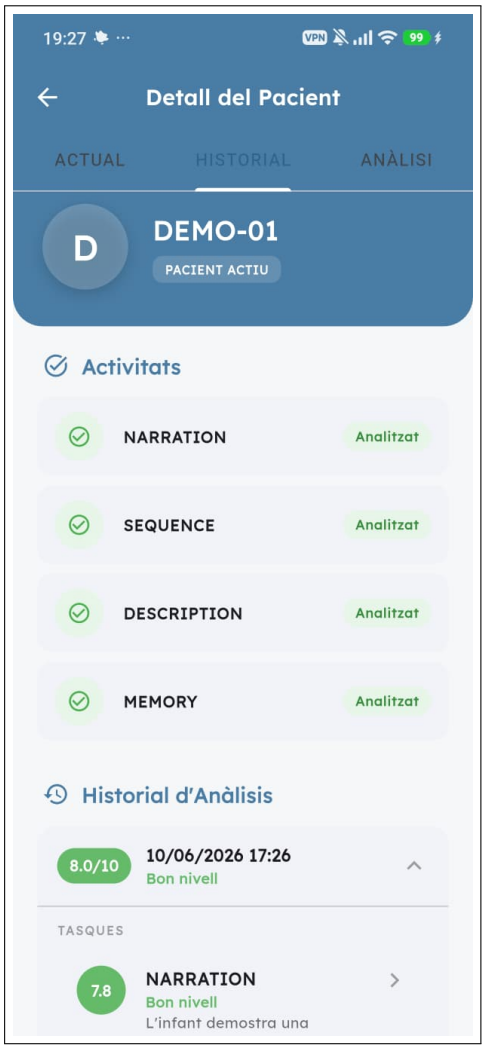
31. Session-level NLP averages and the four global indicators.



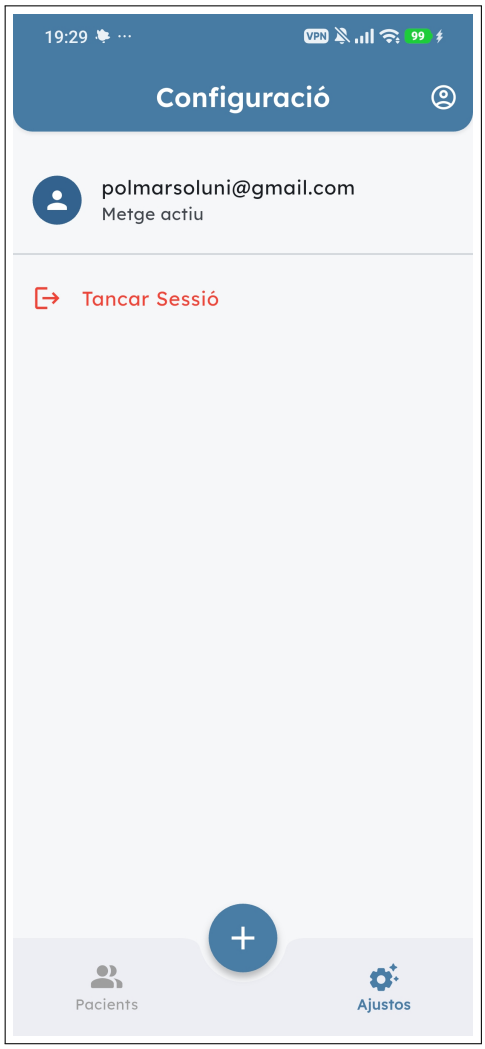
32. Global session report aggregating all tasks with an overall score.



33. Analysis history with per-task scores.



34. Analysed-activities overview.



35. Settings: active account and sign-out.

36. Existing patient details can be edited.

Bibliography

[1] Margaret J Snowling. Early identification and interventions for dyslexia: a contemporary view. *Journal of Research in Special Educational Needs*, 13(1):7–14, 2013. doi: 10.1111/j.1471-3802.2012.01262.x.

[2] Frank R Vellutino, Jack M Fletcher, Margaret J Snowling, and Donna M Scanlon. Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of child psychology and psychiatry*, 45(1):2–40, 2004.

[3] Hollis S Scarborough. Very early language deficits in dyslexic children. *Child Development*, 61(6):1728–1743, 1990. doi: 10.2307/1130834.

[4] Hugh W Catts, Marc E Fey, J Bruce Tomblin, and Xuyang Zhang. A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research*, 45 (6):1142–1157, 2002.

[5] Keith E Stanovich. Explaining the differences between the dyslexic and the garden-variety poor reader: The phonological-core variable-difference model. *Journal of learning disabilities*, 21(10):590–604, 1988.

[6] Maryanne Wolf and Patricia Greig Bowers. The double-deficit hypothesis for the developmental dyslexias. *Journal of educational psychology*, 91(3):415–438, 1999.

[7] Bennett A Shaywitz, Sally E Shaywitz, Kenneth R Pugh, W Einar Mencl, Robert K Fulbright, Pawel Skudlarski, R Todd Constable, Karen E Marchione, Jack M Fletcher, G Reid Lyon, and John C Gore. Disruption of posterior brain systems for reading in children with developmental dyslexia. *Biological Psychiatry*, 52(2):101–110, 2002. doi: 10.1016/S0006-3223(02)01365-3.

[8] Nora Maria Raschle, Jennifer Zuk, and Nadine Gaab. Functional characteristics of developmental dyslexia in left-hemispheric posterior brain regions predate reading onset. *Proceedings of the National Academy of Sciences*, 109(6):2156–2161, 2012.

[9] Fumiko Hoeft, Bruce D. McCandliss, Jessica M. Black, Alexander Gantman, Nahal Zakerani, Charles Hulme, Heikki Lyytinen, Susan Whitfield-Gabrieli, Gary H. Glover, Allan L. Reiss, and John D. E. Gabrieli. Neural systems predicting long-term outcome in dyslexia. *Proceedings of the National Academy of Sciences*, 108(1):361–366, 2011. doi: 10.1073/pnas.1008950108.

[10] Maaike Vandermosten, Jolijn Vanderauwera, Catherine Theys, Astrid De Vos, Sophie Vanvooren, Stefan Sunaert, Jan Wouters, and Pol Ghesquière. A dti tractography study in pre-readers at risk for dyslexia. *Developmental cognitive neuroscience*, 14:8–15, 2015.

[11] Xi Yu, Jennifer Zuk, Meaghan V Perdue, Ola Ozernov-Palchik, Talia Raney, Sara D Beach, Elizabeth S Norton, Yangming Ou, John DE Gabrieli, and Nadine Gaab. Putative protective neural mechanisms in prereaders with a family history of dyslexia who subsequently develop typical reading skills. *Human brain mapping*, 41(10):2827–2845, 2020.

[12] Paula Tallal. Improving language and literacy is a matter of time. *Nature Reviews Neuroscience*, 5(9):721–728, 2004. doi: 10.1038/nrn1499.

[13] Roderick I Nicolson, Angela J Fawcett, and Paul Dean. Developmental dyslexia: the cerebellar deficit hypothesis. *Trends in neurosciences*, 24 (9):508–511, 2001.

[14] Dorothy V. M. Bishop and Margaret J. Snowling. Developmental dyslexia and specific language impairment: same or different? *Psychological Bulletin*, 130(6):858–886, 2004. doi: 10.1037/0033-2909.130.6.858.

[15] Dorothy VM Bishop, Margaret J Snowling, Paul A Thompson, Trisha Greenhalgh, Catalise-2 Consortium, Catherine Adams, Lisa Archibald, Gillian Baird, Ann Bauer, Jude Bellair, et al. Phase 2 of catalise: A multinational and multidisciplinary delphi consensus study of problems with language development: Terminology. *Journal of child psychology and psychiatry*, 58(10):1068–1080, 2017.

[16] David K Dickinson and Allyssa McCabe. Bringing it all together: The multiple origins, skills, and environmental supports of early literacy. *Learning Disabilities Research & Practice*, 16(4):186–202, 2001.

[17] Iluminada Sánchez-Doménech and Beatriz Martín. Games and videogames for dyslexia rehabilitation: neurocognitive and psycholinguistic foundations. *Revista de Educación*, 405:153–176, 2024. doi: 10.4438/1988-592X-RE-2024-405-631.

[18] Nadine Gaab and Yaacov Petscher. Earlybird technical manual, 2021. Technical manual.

[19] Ainara Romero, Urtza Garay, Eneko Tejada, and Arantzazu López De La Serna. Efficacy of berni: A software for preschoolers at risk of dyslexia. *International Journal of Child-Computer Interaction*, 38: 100411, 2023.

[20] Luz Rello. The story behind dytective: how we brought research results on dyslexia and accessibility to spanish public schools. In *Proceedings of the 19th International Web for All Conference*, pages 1–3, 2022.

[21] Monica Berns, Lisa Dieker, and Sherron Roberts. Identifying reading disabilities through eye movements: A validation study using lexplore and ai-driven technology. *Frontiers in Education*, 11:1790777, 2026.

[22] Mercedes Rodrigo, Cristina Rodríguez, Estefanía Rojas, Juan E Jiménez, Remedios Guzmán, Rosario Ortiz, Alicia Díaz, Adelina Estévez, Eduardo García, Isabel Hernández-Valle, et al. Validez discriminante de la batería multimedia sicole-r-primaria para la evaluación de procesos cognitivos asociados a la dislexia. *Revista de Investigación Educativa*, 27(1):49–71, 2009.

[23] Jorge López-Olóriz, Violeta Pina, Sandra Ballesta, Soraya Bordoy, and Laura Pérez-Zapata. Proyecto petit ubinding: método de adquisición y mejora de la lectura en primero de primaria. estudio de eficacia. *Revista de logopedia, foniatría y audiología*, 40(1):12–22, 2020.

[24] Francisca Serrano, JF Bravo Sánchez, and M Gómez Olmedo. Galexia: Evidence-based software for intervention in reading fluency and comprehension. In *Inted2016 proceedings*, pages 2001–2007. IATED, 2016.

[25] Henna Ahmed, Angela Wilson, Natasha Mead, Hannah Noble, Ulla Richardson, Mary A Wolpert, and Usha Goswami. An evaluation of the efficacy of graphogame rime for promoting english phonics knowledge in poor readers. *Frontiers in Education*, 5:132, 2020.

[26] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

[27] Rong Niu, Lu Ni, and Feng Zhu. Emerging technologies and neuroscience-based approaches in dyslexia: a narrative review toward integrative and personalized solutions. *Frontiers in Human Neuroscience*, 19:1683924, 2025.

[28] Erin K Robertson, Catherine Mimeau, and S Hélène Deacon. Do children with developmental dyslexia have syntactic awareness problems once phonological processing and memory are controlled? *Frontiers in Language Sciences*, 3:1388964, 2024.

[29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[30] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[31] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[32] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. `https://doi.org/10.5281/zenodo.1212303`, 2020.

[33] J Martin Bland and Douglas G Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986.

[34] Fiona J Duff, Charles Hulme, Katy Grainger, Samantha J Hardwick, Jeremy NV Miles, and Margaret J Snowling. Reading and language intervention for children at risk of dyslexia: a randomised controlled trial. *Journal of Child Psychology and Psychiatry*, 55(11):1234–1243, 2014.

[35] Raquel Balboa-Castells, Esteban Peñaherrera, Shafaq Rubab, Alfonso Igualada, Mònica Sanz-Torrent, and Llorenç Andreu. Prevenir: An oral language intervention for preventing reading difficulties in kindergarten children—a preliminary effectiveness study. *Language, Speech, and Hearing Services in Schools*, 57(2):462–478, 2026.

[36] Jessica M Lammert, Angela C Roberts, Ken McRae, Laura J Batterink, and Blake E Butler. Early identification of language disorders using natural language processing and machine learning: Challenges and emerging approaches. *Journal of speech, language, and hearing research*, 68(2):705–718, 2025.

[37] Fagun Patel, Duc Quang Nguyen, Sang T Truong, Jody Vaynshtok, Sanmi Koyejo, and Nick Haber. The sound of syntax: Finetuning and comprehensive evaluation of language models for speech pathology. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34907–34925, 2025.

[38] Vibha Tiwari, Ocean Agarwal, Manya Sharma, Rashi Sahu, Radhika Babar, Rebakah Geddam, Muhammad Awais, and Hemant Ghayvat. Akshar mitra: a multimodal integrated framework for early dyslexia detection. *Frontiers in Digital Health*, 7:1726307, 2025.

[39] Ahmed Adel Attia, Jing Liu, Wei Ai, Dorottya Demszky, and Carol Espy-Wilson. Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 74–80, 2024.

[40] Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan. Benchmarking children's asr with supervised and self-supervised speech foundation models. *arXiv preprint arXiv:2406.10507*, 2024.

[41] Cati Riembau, Ignasi Ivern Pascual, and Elisabet Serrat-Sellabona. Phonological processing skills in bilingual (catalan and spanish) students with and without dyslexia. *Revista de Investigación en Logopedia*, 15:159–171, 2025.

[42] Carlos Mena, Pol Serra, Jacobo Romero, Abir Messaoudi, Jose Giraldo, Carme Armentano-Oller, Rodolfo Zevallos, Ivan Meza, and Javier Hernando. Optimizing asr for catalan-spanish code-switching: A comparative analysis of methodologies. *arXiv preprint arXiv:2507.13875*, 2025.

[43] Sameer Pradhan, Ronald Cole, and Wayne Ward. My science tutor (myst)–a large corpus of children's conversational speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12040–12045, 2024.

[44] Carme Armentano-Oller, Montserrat Marimon, and Marta Villegas. Becoming a high-resource language in speech: The catalan case in the common voice corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2142–2148, 2024.