VNiVERSiDAD D SALAMANCA

TRABAJO DE FIN DE MÁSTER

MÁSTER EN INGENIERÍA INFORMÁTICA

FACULTAD DE CIENCIAS

ANEXO 3: Estimación del tamaño y esfuerzo

Sistema de traducción asistida con preservación de composición documental

Computer-assisted translation system with preservation of document layout

Septiembre 2025

Autor: Pablo Caño Pascual

Tutores:

Pablo Chamoso Santos

Guillermo Hernández González

Contenido

1. Introducción

Este documento detalla el proceso de estimación de tamaño y esfuerzo, así como la planificación temporal, que se llevaron a cabo para guiar el desarrollo del proyecto "Sistema de traducción asistida con preservación de composición documental". En cualquier proyecto de ingeniería de software, y en particular en aquellos que integran componentes de investigación y desarrollo, es fundamental realizar un análisis previo que permita establecer un marco de trabajo realista, definir el alcance y anticipar los plazos de ejecución.

La gestión controlada del proyecto se ha apoyado en las métricas y cronogramas que se presentan a continuación. Este documento describe, en primer lugar, la metodología seleccionada para llevar a cabo la estimación de esfuerzo, justificando su idoneidad para el contexto de un Trabajo de Fin de Máster. Posteriormente, se presenta el desglose de los componentes del sistema y la tabla de esfuerzo resultante, que cuantifica el trabajo previsto para cada uno de ellos.

Finalmente, la planificación temporal se visualiza a través de diagramas de Gantt. Se expone tanto la planificación inicial, concebida al inicio del proyecto, como el cronograma real de ejecución. La comparativa entre ambos permite realizar un análisis de las desviaciones, identificando los desafíos encontrados durante el desarrollo y las decisiones tomadas para gestionarlos, ofreciendo así una visión completa del ciclo de vida del proyecto.

2. Metodología de estimación de esfuerzo

Para abordar la estimación de un proyecto con las características del presente TFM, que combina el desarrollo de software tradicional con la investigación y aplicación de tecnologías de vanguardia en inteligencia artificial, se requería una metodología que fuera a la vez estructurada y flexible.

Tras evaluar diferentes enfoques, se descartaron modelos formales como COCOMO II o el Análisis de Puntos de Función (FPA). Si bien son robustos, están diseñados para proyectos de mayor envergadura, desarrollados por equipos y requieren datos históricos que no estaban disponibles en este contexto académico e individual. Del mismo modo, aunque la gestión del proyecto se inspiró en principios ágiles, las métricas relativas como los Puntos de Historia no se traducen directamente en el esfuerzo absoluto (horas-persona) necesario para una planificación temporal formal como la que se requiere en este documento.

En su lugar, se optó por una **Estimación basada en componentes**. Esta metodología híbrida se basa en los siguientes principios:

1. **Descomposición (Divide y vencerás):** El sistema completo se descompone en módulos o componentes funcionales y técnicos manejables.
2. **Estimación ascendente (Bottom-Up):** Se estima el esfuerzo requerido para cada componente individualmente, y la suma de estos esfuerzos parciales conforma la estimación total del proyecto.
3. **Calibración por complejidad:** A cada componente se le asigna un nivel de complejidad, lo que permite asignar un rango de horas más realista, reconociendo que no todas las tareas son iguales.

La elección de esta metodología se vio reforzada por la fase preliminar de Prueba de Concepto (PoC). Los conocimientos adquiridos durante dicha fase permitieron realizar una descomposición del sistema mucho más precisa y una asignación de complejidad más informada, reduciendo la incertidumbre general de la estimación.

El proceso sistemático seguido para obtener las métricas de esfuerzo fue el siguiente:

1. **Descomposición funcional y técnica:** Se identificaron los principales bloques constructivos del sistema, alineados con la arquitectura descrita en el **Anexo 2: Análisis y diseño del sistema**. Se incluyeron tanto tareas de desarrollo (backend, frontend) como tareas transversales (investigación, pruebas, documentación).
2. **Clasificación por complejidad:** A cada componente se le asignó uno de tres niveles de complejidad (Baja, Media, Alta), basándose en criterios predefinidos:
    - **Baja:** Tareas bien definidas, que implican el uso de tecnologías conocidas o librerías estándar con poca personalización. La incertidumbre técnica es mínima.
    - **Media:** Tareas que requieren la integración de varios submódulos, el desarrollo de lógica de negocio específica o la adaptación de herramientas existentes. La incertidumbre es moderada.
    - **Alta:** Tareas con un componente significativo de investigación y desarrollo, alta incertidumbre tecnológica, o que constituyen el núcleo innovador del proyecto.
3. **Estimación de esfuerzo en Horas-Persona:** Se estableció una correspondencia entre el nivel de complejidad y un rango de horas de trabajo estimadas, asumiendo una dedicación individual al proyecto.
    - Complejidad **Baja**: 8 - 16 horas.
    - Complejidad **Media**: 20 - 40 horas.
    - Complejidad **Alta**: 40 - 80 horas.
4. **Aplicación de un factor de contingencia:** Al subtotal de horas estimadas se le añadió un **factor de contingencia del 20%**. Este colchón es fundamental en cualquier proyecto de software para absorber imprevistos,

tareas no contempladas inicialmente, desafíos durante la fase de integración y la resolución de errores, asegurando que la planificación global sea más robusta y realista.

3. Estimación del tamaño y esfuerzo

Siguiendo la metodología descrita en el apartado anterior, el primer paso fue descomponer el proyecto en sus componentes funcionales y técnicos principales. Esta descomposición permite analizar cada parte del sistema de forma aislada para asignarle una complejidad y, consecuentemente, un esfuerzo estimado. A continuación, se presenta el listado de los componentes identificados.

Los componentes en los que se dividió el proyecto, abarcando desde la investigación inicial hasta el despliegue y la documentación final, son los siguientes:

- **C1: Investigación y estado del arte:** Análisis de las tecnologías existentes para segmentación de documentos (Layout Analysis), Reconocimiento Óptico de Caracteres (OCR) y traducción mediante Grandes Modelos de Lenguaje (LLMs).
- **C2: Diseño de la arquitectura del sistema:** Definición de la arquitectura de microservicios, selección de la pila tecnológica (FastAPI, Vue.js, Celery, Redis), y diseño del flujo de datos y comunicación entre contenedores.
- **C3: Desarrollo del backend API (FastAPI):** Implementación de los endpoints REST para la gestión de tareas, carga de ficheros y consulta de resultados. Incluye la validación de datos con Pydantic.
- **C4: Desarrollo del worker asíncrono (Celery):** Creación del pipeline de procesamiento en segundo plano, orquestando las diferentes etapas (segmentación, OCR, traducción y reconstrucción).
- **C5: Integración del módulo de segmentación:** Implementación y encapsulamiento de la lógica para analizar la maquetación de los documentos PDF utilizando modelos de Deep Learning.
- **C6: Integración del módulo OCR:** Desarrollo de la funcionalidad para extraer texto de las regiones de imagen identificadas, utilizando el motor Tesseract.

- **C7: Integración del módulo de traducción:** Conexión con la API externa (OpenRouter), ingeniería de prompts y gestión de las respuestas estructuradas de los LLMs.
- **C8: Desarrollo del módulo de reconstrucción de PDF:** Implementación de la lógica para generar el nuevo documento PDF con el texto traducido y la maquetación preservada, usando la librería ReportLab.
- **C9: Desarrollo del frontend (Vue.js):** Creación de la interfaz de usuario para la carga de documentos, configuración de parámetros y visualización del progreso en tiempo real (polling).
- **C10: Desarrollo del frontend (Vue.js):** Implementación de la vista de resultados, incluyendo la comparación de PDFs y el editor de traducciones interactivo.
- **C11: Infraestructura y despliegue:** Creación de los archivos de configuración de Docker y Docker Compose, y configuración del servidor web (Nginx) como reverse proxy.
- **C12: Pruebas y validación:** Realización de pruebas unitarias y de integración para asegurar el correcto funcionamiento del pipeline completo y la cohesión entre los diferentes servicios.
- **C13: Redacción de la memoria del TFM:** Elaboración del documento escrito del proyecto, incluyendo la investigación, descripción técnica, análisis y conclusiones.

Se puede ver en la Ilustración 1 una tabla en la que se muestran estos componentes junto con su correspondiente complejidad y esfuerzo estimados

ID	Componente	Complejidad estimada	Esfuerzo estimado (Horas)
C1	Investigación y estado del arte	Alta	50
C2	Diseño de la arquitectura del sistema	Media	25
C3	Desarrollo del backend API	Media	30
C4	Desarrollo del worker asíncrono	Alta	65
C5	Integración del módulo de segmentación	Alta	40
C6	Integración del módulo OCR	Baja	8
C7	Integración del módulo de traducción	Media	20
C8	Desarrollo del módulo de reconstrucción de PDF	Media	25
C9	Desarrollo del frontend (carga y progreso)	Media	30
C10	Desarrollo del frontend (resultados y editor)	Alta	40
C11	Infraestructura y despliegue	Media	20
C12	Pruebas y validación	Media	25

C13	Redacción de la memoria del TFM	Alta	70
		Subtotal	448
		Contingencia (20%)	90
		ESFUERZO TOTAL ESTIMADO	**538**

Ilustración 1 Tabla de esfuerzo estimado

La asignación de complejidad se basó en el grado de incertidumbre y el nivel de innovación requerido. Componentes como la **investigación (C1)**, el **worker asíncrono (C4)** y la **integración de segmentación (C5)** se clasificaron como de "Alta" complejidad debido a que constituyen el núcleo innovador del proyecto y requerían la exploración de tecnologías no triviales. Del mismo modo, el **frontend de edición (C10)** y la **memoria (C13)** se consideraron "Altos" por su gran volumen de trabajo.

Otros componentes, como la API del backend (C3) o el módulo de reconstrucción (C8), se estimaron como de complejidad "Media", ya que, aunque requerían un desarrollo sustancial, se basaban en tecnologías y patrones de diseño más estandarizados. Finalmente, la integración del OCR (C6) se consideró "Baja", dado que existen librerías maduras (Pytesseract) que simplifican enormemente esta tarea. El esfuerzo total estimado, incluyendo la contingencia, se fijó en **538 horas**, lo que proporcionó una base sólida para la planificación temporal del proyecto.

4. Planificación temporal

La planificación temporal se ha visualizado mediante diagramas de Gantt, que permiten contrastar la planificación inicial con la ejecución real del proyecto.

1. **La planificación inicial:** concebida al inicio del TFM, establecía un cronograma intensivo con el objetivo de entregar el proyecto en la convocatoria de abril de 2025. Este plan representaba la estimación más optimista con la información disponible en esa etapa. Este plan inicial se puede ver en el diagrama de Gantt de la Ilustración 2.
2. **El cronograma real de ejecución:** refleja el desarrollo efectivo del proyecto, que culminó con la entrega en la convocatoria de septiembre de 2025, se puede ver en la Ilustración 3.

La desviación entre ambos cronogramas no obedece a un retraso, sino a una **decisión estratégica de replanificación** tomada en torno a marzo de 2025. Durante el desarrollo de los componentes de mayor complejidad técnica (C4: Worker asíncrono, C5: Módulo de segmentación), la complejidad real y los desafíos de integración superaron las estimaciones iniciales. Ante esta situación, y en consenso con los tutores, se concluyó que forzar la finalización para marzo comprometería significativamente la calidad y robustez del sistema.

Por tanto, se optó por una extensión controlada del proyecto. Este tiempo adicional permitió abordar adecuadamente los desafíos técnicos, ampliar el alcance funcional del editor de resultados (C10) y llevar a cabo un ciclo de pruebas (C12) mucho más exhaustivo, resultando en un producto final de mayor calidad y mejor documentado. El análisis comparativo de ambos diagramas justifica esta decisión y evidencia la naturaleza iterativa inherente a los proyectos con un fuerte componente de I+D.

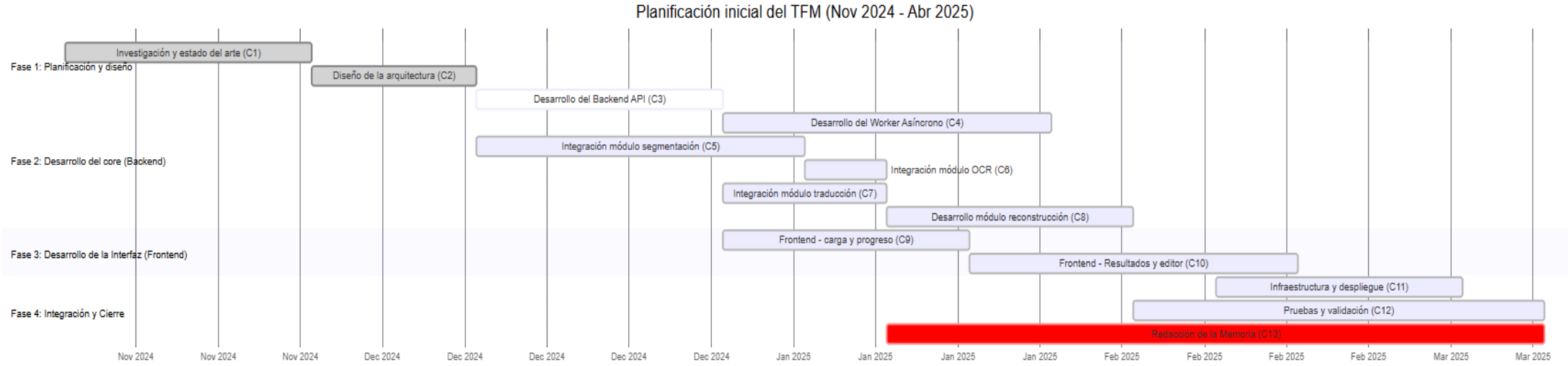

Ilustración 2 Diagrama de Gantt

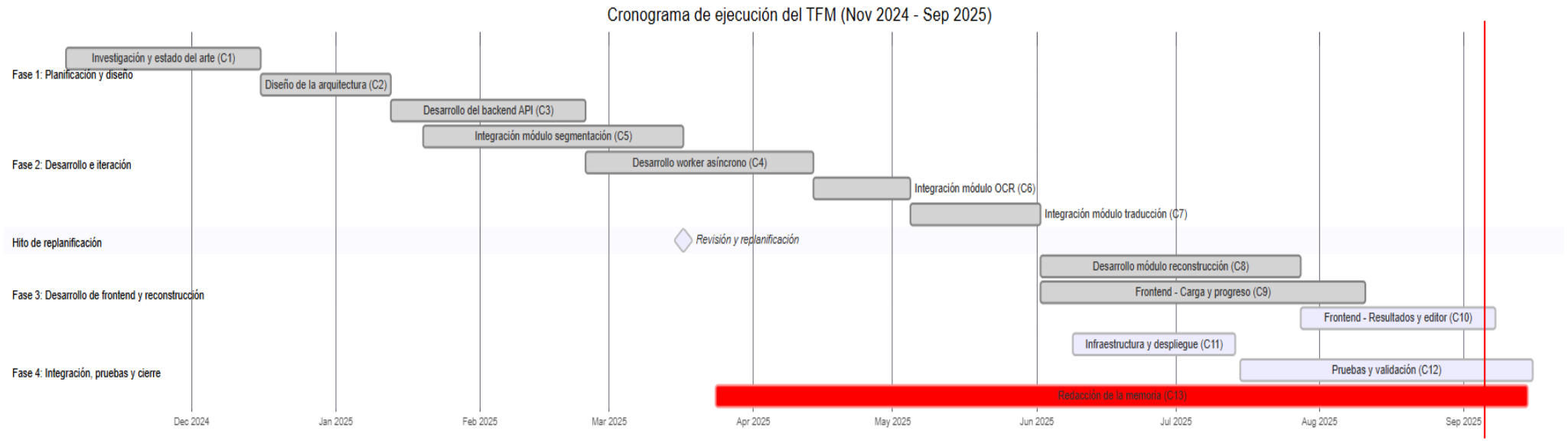

Ilustración 3 Diagrama de Gantt final

Para ofrecer una visión cuantitativa de la ejecución del proyecto, a continuación se presenta una tabla (Ilustración 4) que desglosa el esfuerzo estimado frente al real para cada componente. Los datos permiten evaluar la precisión de la planificación inicial e identificar las desviaciones más significativas.

ID	Componente	Esfuerzo estimado (H)	Esfuerzo real (H)	Desviación (%)
C1	Investigación y estado del arte	50	55	+10%
C2	Diseño de la arquitectura del sistema	25	25	0%
C3	Desarrollo del backend API	30	30	0%
C4	Desarrollo del worker asíncrono	65	70	+8%
C5	Integración del módulo de Segmentación	40	45	+13%
C6	Integración del módulo OCR	8	10	+25%
C7	Integración del módulo de traducción	20	20	0%
C8	Desarrollo del módulo de reconstrucción de PDF	25	28	+12%
C9	Desarrollo del frontend (carga y progreso)	30	38	+27%

C10	Desarrollo del frontend (resultados y editor)	40	55	+38%
C11	Infraestructura y despliegue	20	25	+25%
C12	Pruebas y validación	25	32	+28%
C13	Redacción de la memoria del TFM	70	75	+7%
Subtotal (trabajo base)		448	513	+14.5%
ESFUERZO TOTAL (estimado vs. real)		538	513	-4.6%

Ilustración 4 Comparativa de esfuerzo estimado vs real

El esfuerzo real del trabajo base fue de 513 horas, un 14.5% superior al subtotal estimado. Este incremento se concentra en las áreas donde se encontraron los mayores desafíos: el desarrollo del Frontend interactivo (C9, C10) y las Pruebas (C12).

Es destacable que el esfuerzo real total (513 horas) se mantuvo por debajo del esfuerzo total estimado de 538 horas. Esto demuestra que el factor de contingencia fue efectivo para absorber el sobrecoste derivado de la complejidad imprevista. En conclusión, la extensión del cronograma no se debió a una desviación incontrolada del esfuerzo, sino a una decisión deliberada de redistribuir el trabajo en el tiempo para garantizar un resultado de mayor calidad, gestionando el esfuerzo de manera eficiente dentro de los márgenes previstos.