

Construcción incremental de LEL en español asistida por LLM con trazabilidad e integración semántica

Gabriela Pérez^{1,2}, Joaquina Nogueira¹, Germán Ruiz¹, and Leandro
Antonelli^{1,3}

¹ LIFIA, Facultad de Informática, Universidad Nacional de La Plata

² UNAJ IlyA, Universidad Nacional Arturo Jauretche

³ CAETI, Facultad de Tecnología Informática, Universidad Abierta Interamericana
{gperez, lanto}@lifia.info.unlp.edu.ar
{njoaquina50, geruiz}@gmail.com

Resumen El Léxico Extendido del Lenguaje (LEL) es una técnica utilizada en Ingeniería de Requisitos para capturar el vocabulario específico del dominio y establecer definiciones que faciliten la comprensión del sistema bajo estudio. Sin embargo, su adopción en proyectos reales se ve limitada por la naturaleza tediosa de la construcción manual, la ausencia de trazabilidad entre las definiciones y sus textos de origen, y la escasez de herramientas adecuadas para el español.

Este trabajo presenta un método implementado con LEL-Constructor, una herramienta que aborda estas limitaciones combinando spaCy para la extracción de sintagmas candidatos, un modelo de lenguaje de gran escala (Gemma 3) para la generación y integración de definiciones, y búsqueda semántica sobre embeddings multilingües para detectar entradas redundantes, registrando automáticamente el texto de origen de cada decisión para garantizar trazabilidad. El método mantiene al analista de requisitos como responsable de todas las decisiones. Una evaluación preliminar sugiere que la asistencia automática reduce el esfuerzo inicial en la creación de definiciones sin reemplazar el juicio experto del analista ni su responsabilidad sobre el glosario producido.

Keywords: Ingeniería de Requerimientos · LEL · NLP · Modelos generativos · trazabilidad

1. Introducción

Durante el desarrollo de software, la comunicación entre clientes y desarrolladores constituye un factor clave para el éxito de los proyectos. En las etapas iniciales, suelen surgir dificultades relacionadas con la comprensión del dominio del problema, debido a que ciertos términos adquieren un significado particular en el contexto del sistema, lo que puede generar ambigüedades si no se definen con precisión. La construcción de un glosario que establezca definiciones claras

para estos términos se convierte en una tarea de apoyo fundamental para el proceso de requisitos.

El Léxico Extendido del Lenguaje (LEL) [1] es una técnica utilizada para capturar y documentar el vocabulario específico de un dominio mediante la identificación de símbolos (términos relevantes) y sus definiciones estructuradas en nociones (qué es) e impactos (cómo se comporta o qué efectos produce). Sin embargo, la construcción del LEL presenta varias limitaciones que dificultan su adopción en proyectos reales.

La primera limitación es el costo de la construcción manual. Redactar definiciones estructuradas desde cero demanda tiempo y esfuerzo considerables, lo que genera resistencia a adoptar la técnica en proyectos reales. A medida que el LEL crece con cada nuevo texto, la carga aumenta progresivamente. Sin herramientas de soporte, el analista debe recordar si un término ya ha sido definido bajo otra denominación, lo que conduce a la incorporación de entradas duplicadas o semánticamente solapadas, que afectan la calidad del glosario.

La segunda es la ausencia de trazabilidad. Raramente se registra de manera estructurada el origen de cada término. No se documenta en qué reunión surgió, qué párrafo, quién validó la decisión. Esto impide responder cuando se cuestiona una definición, dificulta la resolución de conflictos entre stakeholders, y elimina la posibilidad de auditar las decisiones tomadas. La tercera es la falta de herramientas para el español. La mayoría de las soluciones existentes son para textos en inglés y no contemplan las particularidades morfosintácticas del español, lo que deja en desventaja a proyectos latinoamericanos, desarrollados en español.

Este trabajo presenta LEL-Constructor, una herramienta que aborda las tres limitaciones mencionadas mediante un flujo de construcción incremental asistida. La herramienta incorpora extracción automática de sintagmas candidatos con spaCy, generación e integración de definiciones con un modelo de lenguaje de gran escala como Gemma 3, y búsqueda semántica basada en embeddings multilingües para la detección de entradas redundantes. La trazabilidad se integra como parte del proceso, registrando automáticamente el texto de origen de cada término y el historial de integraciones. En todo momento, el analista conserva la responsabilidad sobre las decisiones semánticas.

El resto del artículo se estructura de la siguiente manera: la Sección 2 presenta el trabajo relacionado; la Sección 3 describe el método propuesto; la Sección 4 presenta la herramienta LEL-Constructor y su implementación; la Sección 5 ilustra su funcionamiento mediante un ejemplo; la Sección 6 reporta la validación preliminar; la Sección 7 discute los resultados; y la Sección 8 presenta las conclusiones y el trabajo futuro.

2. Trabajos relacionados

La construcción del LEL ha sido abordada desde distintos ángulos en trabajos recientes del área de Ingeniería de Requisitos.

En cuanto a la identificación automática de términos, en [5] proponen un método multicriterio para identificar expresiones candidatas a entradas del LEL

en textos cortos en español. El método combina extracción de sintagmas mediante patrones gramaticales, un sistema de scoring en cinco dimensiones y un modelo de lenguaje local para detectar polisemia. Evaluado sobre tres dominios, alcanza valores de recall entre 0.76 y 0.88. Sin embargo, el trabajo se limita a la identificación de candidatos y no aborda la generación de definiciones ni la detección de entradas redundantes.

En el extremo opuesto, en [3] parten de un LEL ya construido y aplican NLP para extraer conocimiento estructurado en forma de conceptos, atributos y relaciones entre símbolos. Su trabajo evidencia el valor del LEL como fuente de conocimiento estructurado, pero asume que el glosario ya existe y no contribuye a su construcción.

Desde la perspectiva de la integración de LLMs y artefactos de requisitos, en [4] se demuestra que incorporar el LEL como contexto en prompts de LLMs mejora significativamente la precisión terminológica de las user stories generadas a partir de mockups, con puntajes que muestran mejoras en casi todos los tipos de requisitos evaluados. Este resultado subraya la importancia de contar con un LEL de calidad como artefacto previo, motivando directamente el desarrollo de herramientas que asistan su construcción.

LEL-Constructor se diferencia de estos trabajos al abordar el proceso completo de construcción incremental: integra la generación de definiciones asistida por LLM, la detección de redundancias semánticas mediante embeddings multilingües, y la trazabilidad al texto fuente como parte integral del flujo de trabajo. A diferencia de los enfoques mencionados, ninguna definición se incorpora al LEL sin la revisión y aprobación explícita del analista responsable, preservando el control conceptual sobre el glosario producido.

3. Método propuesto

La Figura 1 presenta una visión general del método, que organiza el proceso en cinco pasos secuenciales. El flujo parte de un texto de entrada y avanza desde la identificación de candidatos hasta el registro con trazabilidad, con una decisión explícita del analista en cada etapa. Los pasos se describen a continuación.

3.1. Paso 1: Identificación de Términos Candidatos

El sistema analiza el texto de entrada y extrae automáticamente sintagmas nominales, verbales y de estado que constituyen candidatos potenciales para el LEL. La extracción automática de sintagmas se apoya en la técnica propuesta en [5]. Estos candidatos se presentan al analista en una lista, junto con el fragmento de texto donde aparecen, para que decida cuáles incorporar efectivamente al glosario. Si el analista considera que un término relevante no fue identificado automáticamente, puede agregarlo manualmente. En ambos casos, el término queda vinculado al fragmento de texto que motivó su incorporación, preservando la trazabilidad desde el primer paso del proceso.

Paso 2: Generación y Verificación Simultánea

Una vez que el analista selecciona un término candidato, el sistema ejecuta dos acciones en paralelo: la generación asistida de una definición y la búsqueda de entradas similares en el LEL existente.

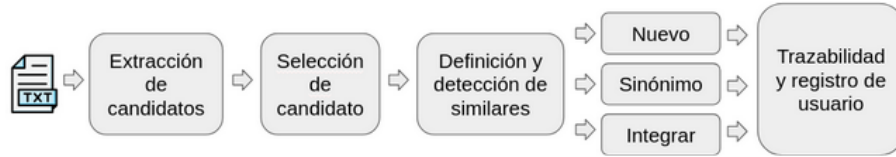


Figura 1. Visión general del método de construcción incremental del LEL.

La definición se genera a partir del término seleccionado y el fragmento de texto de origen. El LLM produce una propuesta estructurada según el formato LEL estándar: noción (qué es el término) e impactos (cómo se comporta o qué efectos produce en el dominio). La propuesta se presenta al analista responsable, quien puede aprobarla o editarla manualmente antes de incorporarla al glosario.

En paralelo, el sistema calcula el embedding vectorial del término y realiza una búsqueda por similitud coseno contra todas las entradas ya almacenadas en el LEL, presentando las tres más similares junto con sus puntajes de similitud y sus definiciones completas.

El analista visualiza simultáneamente la definición propuesta y las entradas similares del LEL existente, lo que le permite tomar una decisión informada en un único paso.

3.3. Paso 3: Decisión del analista

Con base en la información presentada, el analista dispone de tres opciones:

Término NUEVO: el término representa un concepto distinto de los similares encontrados. El analista edita la definición propuesta por el LLM y la aprueba. La entrada se almacena en el LEL junto con su referencia al texto de origen.

SINÓNIMO: el término ya está cubierto por una entrada existente bajo otra denominación. El analista descarta el nuevo término, opcionalmente registrando el sinónimo como referencia cruzada en la entrada existente.

INTEGRAR: el término aporta información nueva que debe enriquecer una entrada ya existente. El analista selecciona la entrada con la que debe integrarse e inicia el proceso descrito en el Paso 4.

3.4. Paso 4: Integración Asistida

Cuando el analista decide integrar el término candidato con una entrada existente, el sistema invoca al LLM proporcionándole tanto los textos de origen de ambas entradas como sus definiciones actuales. El LLM propone una entrada fusionada que es fiel a ambos textos fuente, combinando nociones e impactos sin perder el contexto original de ninguna de las dos.

El analista revisa la propuesta, la edita si lo considera necesario y la aprueba. La entrada integrada reemplaza a la anterior en el LEL, manteniendo trazabilidad a ambos textos de origen.

3.5. Paso 5: Trazabilidad y Registro

Toda decisión queda registrada automáticamente. Para cada entrada del LEL se almacena: el fragmento de texto que motivó su incorporación, el analista responsable de la validación, la fecha y hora de la acción, y en caso de integración, las referencias a ambos textos de origen y la estrategia aplicada.

Esta trazabilidad opera de forma activa: cuando el cliente cuestiona una definición, el analista puede señalar con precisión el documento y el párrafo que la originaron; cuando surge un conflicto entre stakeholders, el historial permite reconstruir el razonamiento que llevó a la definición actual.

4. LEL-Constructor

LEL-Constructor es un prototipo que implementa el método propuesto y está orientado a analistas de requisitos que utilizan LELs en sus proyectos, y que necesiten construir o ampliar un glosario de dominio a partir de notas o resúmenes de reuniones. La herramienta asiste en cada etapa del proceso sin reemplazar el criterio del profesional. Genera definiciones, detecta redundancias y sugiere integraciones, que solo se incorporan al LEL con la aprobación explícita del responsable.

Está implementado como una aplicación web local en Python con arquitectura cliente-servidor. El frontend en HTML, CSS y JavaScript se comunica con un backend FastAPI mediante llamadas REST, permitiendo una interfaz liviana mientras el procesamiento se realiza localmente, sin depender de servidores externos.

El backend se organiza en tres módulos principales. El módulo de extracción utiliza spaCy con el modelo `es_core_news_lg` para identificar sintagmas candidatos mediante análisis morfosintáctico del texto de entrada. El módulo de generación gestiona la comunicación con Gemma 3 a través de Ollama, generando una definición estructurada de la entrada LEL a partir del término y su texto de origen. La elección de Gemma 3 (version 12B) como modelo de lenguaje se basó en criterios de viabilidad ya que es un modelo de código abierto que puede ejecutarse localmente, sin requerir hardware especializado.

El módulo semántico calcula el embedding vectorial de cada término usando paraphrase-multilingual-mpnet-base-v2 de Sentence Transformers [2] [6] y realiza búsqueda por similitud coseno contra todas las entradas existentes, retornando las tres más similares junto con sus puntajes.

Todas las dependencias se ejecutan localmente, garantizando que los textos procesados no abandonen la máquina del analista en ningún momento.

Ejemplo y funcionalidad

Para ilustrar el funcionamiento de la herramienta, se presenta un ejemplo en el dominio de controladores de tráfico aéreo a partir del siguiente fragmento:

”El papel de los controladores de tráfico aéreo es más esencial que nunca. A medida que aumenta el volumen de aeronaves y los sistemas de control de tráfico aéreo civil adoptan la transformación digital, los controladores deben adaptarse a nuevos desafíos. Son el elemento humano que garantiza la supervisión del tráfico aéreo en todas las condiciones tanto meteorológicas, como fallos técnicos o amenazas a la seguridad.”

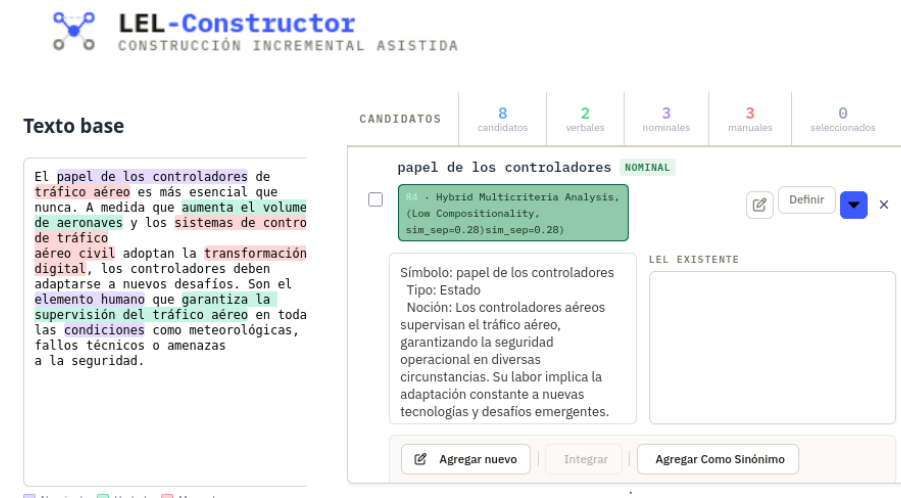


Figura 2. LEL-Constructor: prototipo de la herramienta de soporte para la construcción incremental de LEL.

La Figura 2 muestra el prototipo durante el proceso de identificación de candidatos. El analista selecciona el término "elemento humano". El LLM genera la siguiente propuesta:

Símbolo: elemento humano

Tipo: Sujeto

Noción: Persona que garantiza la supervisión del tráfico aéreo en diversas condiciones.

Impactos:

- Los controladores de tráfico aéreo son el elemento humano que supervisa el tráfico aéreo.
- El elemento humano debe adaptarse a nuevos desafíos derivados del aumento del volumen de aeronaves y la transformación digital de los sistemas de control.
- El elemento humano es responsable de la supervisión en condiciones meteorológicas, fallos técnicos o amenazas a la seguridad.

La búsqueda semántica no encuentra entradas similares, ya que el LEL se encuentra vacío. El analista aprueba la definición y el término queda incorporado al LEL con referencia al fragmento de origen — caso Nuevo.

A medida que el LEL crece, la búsqueda semántica permite detectar términos redundantes. Si al analizar un candidato se detecta una entrada existente con alta similitud, el analista puede optar por registrarlo como Sinónimo, o por Integrar ambas definiciones en una entrada fusionada asistida por el LLM.

6. Validación preliminar

El prototipo fue utilizado por cinco estudiantes de la Licenciatura en Sistemas, con conocimientos de la técnica LEL. Al finalizar, compartieron sus observaciones. Si bien el alcance de esta validación no permite extraer conclusiones estadísticamente significativas, los comentarios son consistentes con los objetivos del enfoque y aportan evidencia cualitativa sobre su utilidad, tales como:

- Utilidad de la búsqueda semántica. Valoraron positivamente la presentación de entradas similares durante la incorporación de un nuevo término. Señalaron que esta funcionalidad permite reconocer si un candidato representa el mismo concepto que una entrada existente, una tarea que se vuelve más compleja de hacer manualmente a medida que el LEL crece.
- Proceso revisar–ajustar–aprobar. Los participantes consideraron natural el proceso propuesto. La posibilidad de editar la definición generada por el LLM antes de aprobarla fue destacada como un aspecto positivo, ya que permite mantener el control sobre el contenido.
- Riesgo de alucinaciones del LLM. Señalaron la preocupación de que el modelo incorpore información incorrecta o no presente en el texto de origen. Esta observación refuerza una decisión de diseño central del método, que ninguna definición se incorpora al LEL sin la revisión y aprobación explícita del responsable.

7. Discusión

Los resultados de esta validación exploratoria son consistentes con el objetivo principal del enfoque: reducir el esfuerzo de construcción del LEL sin desplazar el criterio del profesional responsable. Sin embargo, es importante señalar que la preferencia reportada por la herramienta frente a la construcción manual no constituye evidencia de que el LEL producido sea de mayor calidad; únicamente indica que el proceso resultó menos costoso en términos de esfuerzo inicial.

La preocupación señalada sobre alucinaciones del LLM es esperable en cualquier sistema de generación asistida y no invalida el enfoque, al contrario, refuerza la pertinencia del diseño centrado en la revisión humana. La combinación de un modelo que propone definiciones y un analista que las valida constituye una alternativa más robusta que cualquiera de los dos actuando de manera independiente. El modelo reduce el esfuerzo de redacción y el analista garantiza la coherencia.

Conclusiones y trabajo futuro

Este trabajo presenta LEL-Constructor, un prototipo que implementa un método de construcción asistido e incremental del LEL en español. El método articula tres tecnologías complementarias, como la extracción de sintagmas con spaCy, generación de definiciones con un modelo de lenguaje de gran escala y búsqueda semántica basada en embeddings, para asistir en una tarea que históricamente demanda un esfuerzo considerable y es propensa a inconsistencias.

El aporte del enfoque reside en dos decisiones. La primera es mantener al analista de requisitos como responsable de todas las decisiones semánticas. El sistema propone, detecta similitudes y sugiere integraciones, pero ninguna definición se incorpora al LEL sin su aprobación explícita. La segunda es incorporar la trazabilidad como parte integral del proceso, registrando automáticamente el texto de origen de cada término, el responsable de la decisión y el historial de integraciones, lo que transforma al LEL en un artefacto auditable.

La validación preliminar sugiere que la asistencia automática reduce el esfuerzo inicial sin desplazar el criterio profesional. Sin embargo, el alcance es limitado y no permite afirmar que el LEL producido sea de mayor calidad que uno construido manualmente.

Como trabajo futuro, se planea una evaluación controlada con más participantes y métricas cuantitativas de calidad y consistencia. La integración con [5] completaría un pipeline totalmente asistido desde las notas de reunión hasta el LEL consolidado. Adicionalmente, se prevé evaluar la herramienta mediante SUS o TAM, con el objetivo de obtener evidencia sobre la usabilidad y adopción de LEL-Constructor por parte de analistas de requisitos.

Referencias

1. Leite, J.C.S.P., Franco, A.P.M.: A Strategy for Conceptual Model Acquisition. In: Proceedings of the 1st IEEE International Symposium on Requirements Engineering, pp. 243–246. IEEE Computer Society Press, San Diego (1993).
2. Reimers, N., Gurevych, I.: Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. EMNLP 2020.
3. L. Antonelli, M. Lezoche, and J. Delle Ville, “Knowledge Extraction from the LEL Glossary Using Natural Language Processing,” *Tecnológicas*, vol. 27, no. 59, e2917, Apr. 2024. <https://doi.org/10.22430/22565337.2917>
4. Firmenich, D., Antonelli, L., Pazos, B., Lozada, F., Morales, L.: Exploring LLMs for User Story Extraction from Mockups. In: Proceedings of WER 2025, Rio de Janeiro, Brazil (2025).
5. Pérez, G., Mostaccio, C., Maltempo, G., Bogado, J., Antonelli, L. Automatic identification of domain expressions to build a LEL glossary from Spanish texts. Extended Abstract. 3rd Workshop on Multi-disciplinary, Open, and integRatEd Requirements Engineering (MO2RE), ICSE 2026, Rio de Janeiro, Brazil, April 12, 2026.
6. Hugging Face, <https://huggingface.co/>, Accedido en marzo 2026.