Creación de un lexicón biomédico en español por emulación sistémica para su uso en un sistema de información en salud

Clusella M.[1], Luna P.D.[1], Mitre M.[1], Gómez A.[2], Martínez M.[2], Luna D.[2], Lopez Osornio A.[2], Gonzalez Bernaldo de Quirós F.[2], Otero P.[2], Avila H.[3], Luna P.A.[3]

[1]Colegio Mayor Universitario Santiago del Estero, Grupo de Ingeniería de Software, Santiago del Estero, Argentina
[2]Departamento de Información Hospitalaria, Hospital Italiano de Buenos Aires, Argentina
[3]Prospectivar Consultores S&M, Santiago del Estero y Buenos Aires, Argentina

Resumen

En el contexto de la creación de un servidor de terminología clínica para ser utilizado por los sistemas de información del Hospital Italiano de Buenos Aires se decidió la creación de un lexicón biomédico en español. El mismo servirá como herramienta lexical, para el procesamiento ortográfico, morfológico y sintáctico de textos libres, en diferentes dominios tales como el ingreso de datos en la lista de problemas de la historia clínica electrónica, la codificación automática del informe de hospitalización, las interfaces de búsqueda para la solicitud de exámenes complementarios y prescripciones farmacológicas entre otros.
Mediante un procedimiento de emulación sistémica Input-Precess-Output de un sistema en inglés de actual vigencia, el SPECILIST lexicon del UMLS, se dirige a la construcción de un Lexicón computacional, elaborado con el objeto de servir de soporte representacional a diversas aplicaciones en el ámbito de las tecnologías del procesamiento del lenguaje humano o lenguaje natural, que procesa el español oficial y el vocabulario biomédico mas usado en el ámbito de la organización Hospital Italiano BA. El estado actual del proyecto se encuentra en el final del diseño, avanzado el desarrollo e implementación y prácticamente lista la corrida de prueba del prototipo Beta.

Palabras Clave
Lexicón, Tesauro, Procesamiento de lenguaje natural, Herramientas lexicales, Codificación semántica.

Introducción

Las últimas décadas han mostrado múltiples avances en la investigación y desarrollo de nomenclaturas para la representación del conocimiento en el campo de la salud. Dejando de manifiesto las dificultades que se presentan en este ámbito [1]. Entre los principales objetivos de las terminologías clínicas se encuentran: la indización de documentos, el soporte para la toma de decisiones, el control y monitoreo de órdenes médicas, técnicas de inferencia, etc. [2]. Dicha información clínica necesita ser expresada usando un comprehensivo y estructurado sistema de terminología.
Pueden diferenciarse tres tipos de terminologías que deben ser manejadas conjuntamente en los sistemas de información en salud [3]:

- *Terminología de interface o aplicación:* es la que debe interactuar con el usuario, debe contener términos y sinonimia de la jerga local.
- *Terminología de referencia:* representada en una compleja base de conocimiento rigurosamente controlada con reglas y relaciones entre los conceptos.
- *Terminología administrativa:* también denominados set de codificación, es una colección de expresiones codificadas usadas en el ámbito financiero y de gestión en las instituciones de salud.

Si bien aún no esta generalizado el uso estandarizado de servicios de terminología [4] existen varios antecedentes con respecto a la definición de dicho servicio. Entre los cuales se encuentran las especificaciones del "Terminology Query Services (TQS)" [5], las del "Terminology Query Language (TQL)" [6] o las del "Common Terminology Services (CTS)" [7].
Un **servidor de terminología** es entendido como una aplicación que agrupa las tres capas previamente descriptas y está compuesto por dos componentes básicos, por un lado una representación conceptual del conocimiento (conceptos) y por el otro herramientas para designación lingüística (términos) [4, 8, 9]. Mientras que la representación conceptual es esencial para la clasificación y manipulación informática de los conceptos, la designación lingüística es crucial para tratar con el lenguaje utilizado cotidianamente. Uno de los mejores exponentes que ha mostrado capacidad para el manejo léxical es el SPECILIST lexicon [10, 11] del Unified Medical Language System (UMLS) [12] por ese motivo se decidió la emulación sistémica de dicho aplicativo.

El propósito central de esta comunicación es informar sobre la experiencia metodológica generada en las distintas etapas de la construcción por emulación sistémica de un lexicón biomédico en español para su utilización en el sistema de información del Hospital Italiano de Buenos Aires (HIBA).

Sistema de información del Hospital Italiano

El *Hospital Italiano de Buenos Aires* es un hospital universitario de alta complejidad que cuenta con 550 camas de internación y más de 400 consultorios ambulatorios en 20 centros de atención distribuidos en Capital Federal y el Gran Buenos Aires. Desde el año 1998 la institución está llevando a cabo una reingeniería de sus sistemas para posibilitar la integración de múltiples fuentes de información existentes (área administrativa, laboratorio, diagnóstico por imágenes, etc.) [13] junto a un proceso de informatización del registro médico con la creación de una historia clínica electrónica (HCE). El desarrollo de dicho sistema tiene como objetivo brindar soporte en la faz asistencial, docente, científica y de gestión, de allí surge el proyecto denominado **"Itálica"** [14, 15]. En el nivel ambulatorio de atención se creó una HCE centrada en el paciente y orientada a problemas, en donde los médicos, durante el proceso asistencial, cargan mediante texto libre narrativo los problemas de salud en dicho sistema. Posteriormente dicha información es codificada secundariamente en forma centralizada por un grupo de codificadores mediante un proceso de calidad auditada [16]. Al realizar un proceso de limpieza y mejora continúa de la base de información, conformada por los problemas ingresados por los profesionales, se desarrolló un sistema de autocodificación de dicha lista de problemas [17]. Posteriormente se realizó

la codificación múltiple de dichos problemas con CIAP 2, CIE-10 y SNOMED CT [18]. En el ámbito de la internación se codifican los diferentes componentes del informe de hospitalización con CIE-9-CM y GRD para la medición de case mix y gestión [19]. Por otro lado se crearon terminologías de interface para dominios tales como: solicitud de exámenes complementarios [20], prescripción de fármacos [21], material descartable y dispositivos médicos [22].
Si bien Itálica ya contiene sistemas de soporte clínico para la toma de decisiones, tales como el sistema notificador de interacciones farmacológicas [23, 24], se decidió avanzar en la creación de un monitor de eventos clínicos para la incorporación de recordatorios y alarmas en la HCE. En paralelo a lo antedicho también se decidió la creación e implementación de un servidor de terminología que permitiera la integración de los múltiples vocabularios utilizados en el sistema de información. En este contexto se decidió la creación de un lexicón en español para ser utilizado por Itálica denominado Lexicón HIBA.

Enmarque referencial

Para la comprensión del dominio técnico es necesario exponer los marcos referenciales enmarcados estos en la *sistémica*, la *informática* y la *computación*, para presentar las metodologías utilizadas en el desarrollo del proyecto del Lexicón HIBA. Ante el requerimiento específico de emulación del SPECIALIST lexicon del UMLS [25], la sistémica fue la opción metodológica empleada para obtener, a través de modelizaciones recursivas de los procesos tipo caja negra, *Input-Process-Output* (IPO), los distintos modelos hipotéticos. Los esfuerzos realizados, se centraron en un principio, en un proceso de aprendizaje/investigación sobre el material bibliográfico disponible para comprender el sistema emulado. Recursivamente se obtuvieron seis niveles de desagregación en los cuales se comprenden las herramientas léxicas constitutivas del SPECIALIST lexicon y sus procesos.

Mediante la emulación sistémica, entendida como la réplica análoga del sistema real, manejada desde la máquina abstracta a través de un razonamiento lógico-deductivo y mediante las modelizaciones que fueron necesarias realizar se obtuvieron tres modelos hipotéticos emulantes:

- *Modelo Hipotético 1:* fue el primer modelo emulante obtenido con el que se descubrió la especificidad para conocer por analogía como funciona el Lexicon en inglés.
- *Modelo Hipotético 2:* es el emulante en español, la meta buscada. Presentado con los diagramas E/R correspondientes, acompañados por la descripción de la estructura de los datos y las especificaciones necesarias para poderlos convertir en lenguaje de máquina Java. Se considera a este emulante, como el modelo meta, un modelo complejo, al que se debe alcanzar a través de los refinamientos recursivos necesarios de la prototipación del Modelo Hipótesis 3.
- *Modelo Hipotético 3:* definido como un modelo operante, por ser más simple en su estructura y en su funcionalidad. Sus corridas de prueba por refinamiento recursivo, al tiempo que prueba las reglas gramaticales, generan los registros que surgen desde la base empírica de la lista de problemas de Itálica desde el año 1998.

Debido a la expansión de la capacidad interactiva de los profesionales, el diseño y desarrollo de interfaces de usuario fueron realizados adecuadamente mediante la adopción de paradigmas y entornos simbionómicos. Los diseños de Ingeniería de Software Orientada a Objetos (ISOO) siguen al paradigma centrado en diversos tipos de usuarios, al estilo y sugerencia del modelo Computer Human Interaction (CHI) de la *"Asociation for Computer Machinery (ACM)"* [26] y los lineamientos de la *"Asociación Interacción Persona Ordenador, (AIPO)"* [27] para desarrollos específicos en la ingeniería de software. Estos lineamientos permiten obtener mayor usabilidad, accesibilidad y satisfacción de los sistemas iteractivos que se desarrollan a través de mejores diseños de interfaces de usuario. No se deberian descuidar –sostienen estos antecedentes- los contextos en el enfoque de intervinculación e interacción entre usuarios (profesionales y organizaciones) y asistentes software. Los diseños en detalle fueron realizados solo para el modelo hipotético 3, por ser la base para el *prototipo beta*, cuyo objetivo principal es el de probar las reglas gramaticales desarrolladas, que permitan realizar el análisis sintáctico y morfológico correspondiente a cada entrada representada en la lista de problemas médicos generados por Itálica que forman la base empírica del proyecto.
Los diseños ISOO fueron plasmados con la técnica Standard UML-UP [28] necesaria para la interrelación y supervisación entre los equipos intervinientes. Por ser un proyecto colaborativo entre distintos equipos técnico-profesional de alcance multidisciplinario, el desafío resulto satisfactorio. Para el *prototipo beta* se llegó al nivel de especificaciones (de cada uno de los objetos, atributos, tablas y bases intervinientes en el modelo hipotético 3) lógicamente sentenciales para probar la coherencia de las reglas y si son validables. A partir de la experticia lograda por el procesamiento semiautomático de los problemas médicos, que delimitaron los requerimientos, se establecieron los requisitos de ingeniería de software para concretar la funcionalidad del sistema de manera definitiva. Exigencia formal de la ingeniería, cuya previsión permite la expansión o evolución del sistema, para lo cual deben estar en correspondencia, casi como par ordenado con los requerimientos del usuario y luego corrobados en la prototipación.

Modelos Emulantes

Es necesario para la acabada comprensión de la descripción del diseño entrar en algunos aspectos de detalle. La modelización sistémica y en particular la emulación conjunta con prototipación exigen un esfuerzo de aproximación de orden metodológico.

Sistema "emulado", descripción sistémica

Durante la fase inicial del proyecto, luego de revisar bibliografía existente sobre el SPECIALIST lexicon [11, 25, 29, 30], el equipo de trabajo estableció como meta la "emulación sistémica" (modelo caja negra/ modelo I-P-O) de un sistema de probada capacidad y utilización extendida, pero para el idioma inglés. El *Unified Medical Language System Knowledge Sources* (UMLS KS) fue utilizado para interpretar y refinar las entradas del usuario, para mapear los términos de usuario con un apropiado vocabulario controlado ya que contiene esquemas de clasificación para interpretar el lenguaje natural. Dicha fuente permite a los profesionales de la salud y a los investigadores el uso de información disponible automáticamente y ayuda al desarrollo de interfaces efectivas que

asistan al usuario. Las bases de conocimiento del UMLS facilitan el desarrollo de aplicaciones de indexación. Tres son las grandes fuentes encontradas en este primer nivel (Nivel 0) de desagregación:

- El ***Metathesaurus*:** es una base de datos de información sobre los conceptos que aparecen en el campo de la biomedicina. Su organización está orientada a conceptos. Contiene información semántica sobre los conceptos biomédicos, sus diferentes nombres y las relaciones entre ellos. Contiene e interconecta muchos vocabularios biomédicos estandar. Es el vocabulario central del UMLS. La utilidad del Metatesauro se aumenta o incrementa cuando es usado en combinación con el SPECIALIST lexicon, con sus programas léxicos y con la red semántica. La edición 2003AB contiene 900.551 conceptos; 2,500 millones de nombres de Conceptos en su vocabulario fuente, y el 64.6 % de los conceptos no tienen restricciones adicionales.
- La ***Semantic Network***: es una red de categorías generales (ontología) o tipos semánticos con los que han sido asignados todos los conceptos en el Metathesaurus.
- El ***SPECIALIST lexicon***: contiene información sintáctica sobre términos biomédicos y eventualmente cubriría la mayoría de los términos componentes en los nombres de los conceptos presentados en el Metathesaurus. Se define sobre un dominio.

Luego del estudio profundo del sistema emulado por aproximación sistémica, se llegó a un nivel de desagregación que permitió entender el modelo caja negra. Definido el Input, como términos, palabras, texto en inglés biomédico, y el output como registros tratados lexicalmente se establecieron los programas léxicos primarios utilizados por el lexicon:

- Normalizador (Norm)
- Generador de Índice de palabras (WordInd)
- Generador de variables léxicas (LVG)

Las mismas son herramientas poderosas para la búsqueda, indexación y el procesamiento léxico. El SPECIALIST lexicon con sus programas léxicos primarios proveen la información léxica necesaria para un sistema de Procesamiento de Lenguaje Natural (NLP). A través de ellos, para cada palabra o término que es ingresado en el lexicón, se registra la información sintáctica, morfológica y ortográfica necesitada por el NLP. Las herramientas léxicas son diseñadas con el objetivo de direccionar el alto grado de variabilidad de las palabras o términos del lenguaje natural. La Figura 1 explica el modelo I-P-O de este segundo nivel de desagregación (Nivel 1).

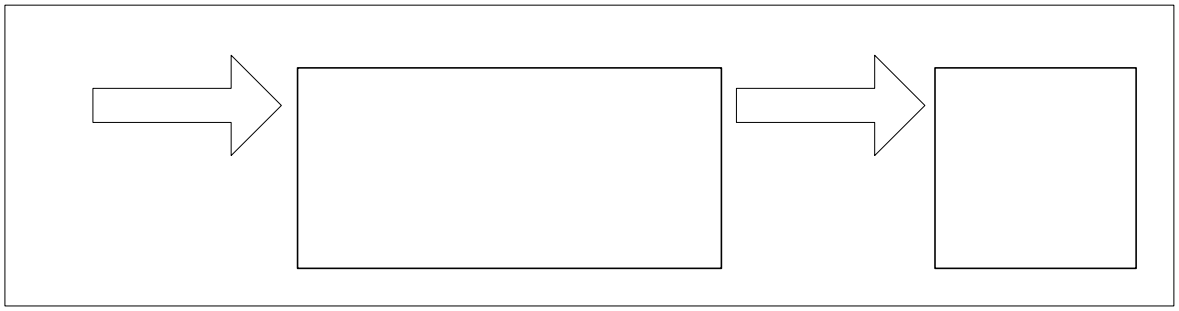

Figura 1: Nivel 1 de desagregación

Luego se avanzó a otro nivel de emulación sistémica de los procesos involucrados en los programas léxicos, en donde se determinaron los respectivos input y output. Como las palabras a menudo tienen formas flexivas, las cuales deben ser consideradas instancias de una misma palabra, el lexicón a través de sus programas léxicos primarios, determina un conjunto de variantes léxicas por cada variante de una entrada o un conjunto de variantes de una parte particular del discurso. La gran ventaja de las herramientas léxicas es que permiten al usuario abstraerse de este tipo de variaciones.

El programa léxico **Norm** (Normalizador) genera cadenas de texto normalizadas que son usadas en el índice de "strings normalizados" (MRXNS). La normalización involucra los procesos de remover genitivos, reemplazar signos de puntuación por espacios, remover fin de palabra, convertir a minúscula cada palabra, separar el string en sus palabra constitutivas y ordenar alfabéticamente las palabras. Quita la inflexión a cada palabra si esta aparece en el lexicón y de no ser así la genera algorítmicamente. Cuando una forma puede ser una inflexión de más de una forma base, se realiza un nuevo proceso de normalización, retornando múltiples formas desinflexionadas. El registro de salida de Norm incluye todos los campos del registro de entrada y agrega un campo adicional con la forma normalizada del string de entrada.

El programa léxico **WordInd** (Word Index), separa los strings en palabras para usarlo con el indexador de palabras MRXNW. La palabra en este caso, es definida como un **token** que contiene más de un carácter alfanumérico separado por espacio o un signo de puntuación. Este programa tokeniza, es decir identifica elementos básicos del lenguaje como palabras, números, símbolos.

El tercer programa léxico es el **LVG** (Generador de variantes léxicas). Este programa genera variantes de las palabras del Input. Consiste de varios componentes de flujo diferentes que pueden ser combinados de distintos modos para producir las variantes léxicas. Determina la categoría sintáctica y la información de inflexión de la entrada. Para el output este programa genera registros de salida para cada variante generada y agrega cinco nuevos campos a un registro de entrada. Los primeros tres campos son idénticos al input, el resto es reemplazado por la información del LVG. El primer campo adicional es la variante generada. El segundo campo adicional corresponde a la categoría sintáctica codificada de la variante. El tercero corresponde a la inflexión de la variante, el cuarto campo adicional indica el flujo que se seleccionó, y el quinto es el número de flujos que generó la variante. La figura 2 sintetiza el proceso desagregado en el Nivel 2.

Los procesos recursivos de la modelización sistémica permitieron avanzar en la desagregación de niveles de entendimiento de la caja negra, a medida que el proceso se comprendía. Así se llegó a un nivel 3 donde se comprendió la estructura y propiedades de los datos de entrada, de las tablas intervinientes (tablas relacionales y auxiliares), de las bases de datos utilizadas y los flujos involucrados entre los distintos programas léxicos primarios para la obtención del registro único de salida con la información léxica codificada. (Ver detalles en Figura 3). Una vez determinados y comprendidos los atributos de las tablas intervinientes pudo vincularse funcionalmente el nivel estructural (Nivel 4), para la obtención del output. Ya en este nivel de desagregación se llegó a las especificaciones analíticas (Nivel 5), se elaboró el manejo de interfaces de carga de

ingresos de los Input (vinculación con Itálica), vinculaciones con otros sistemas del HIBA, conexiones con UMLS-KS, SNOMED CT, etc. De esta manera se obtuvo el Modelo "existente" a "emular" sistémicamente que constituye Modelo Hipótesis 1 para ser el primer "emulante" en español.

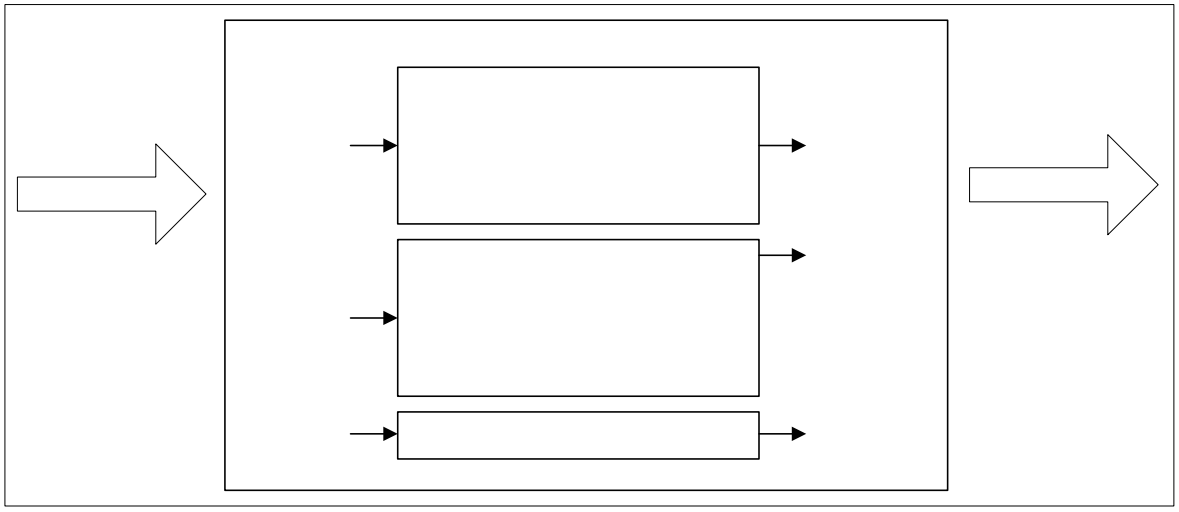

Figura 2: Nivel 2 de desagregación

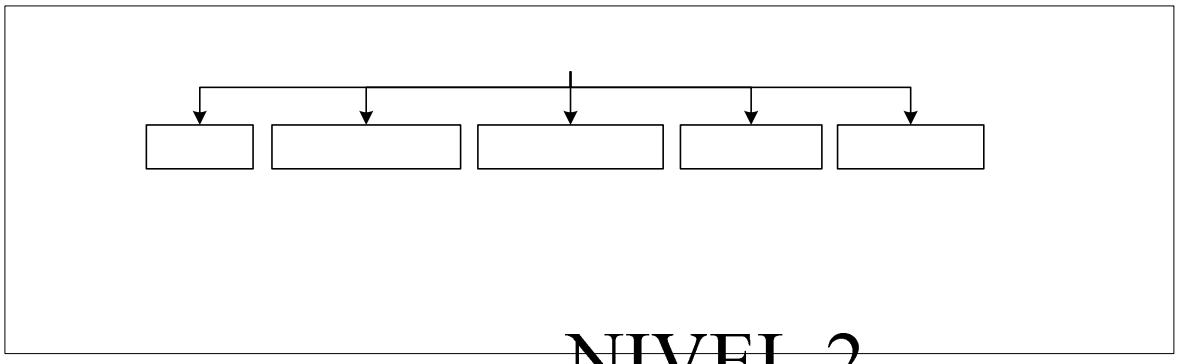

Figura 3: Nivel 3 de desagregación

Todo modelo es único en tanto representación de un sistema, aunque no es el único posible de elaborar; la multiplicidad de enfoques, de perspectivas, de puntos de vista, de intereses, en fin, de contextos y entornos orientan hacia la riqueza de la innovación y de la creatividad. De esa manera se llegaron a planear tres modelos-hipótesis que delimitan el proceso.

Modelo hipotético # 1

El *modelo hipotético 1*, es el modelo del sistema emulado en inglés comprendido desde la metodología de la modelización sistémica empleada para el estudio e investigación del sistema en inglés como existente. Tiene representación gráfica en la Figura 4, ya que es una de las formas en que un sistema puede ser representado. El diagrama E/R (Entidad/Relación) refleja el flujo de los datos que vincula funcionalmente a las tablas intervinientes en el proceso léxico que se genera desde el input, determinada por un

Palabras
Términos
Textos

INPUT

Palabras

Términos

Textos

aplicativo a raíz de la cual intervendrán los programas léxicos primarios correspondientes para realizar sus funciones específicas y así obtener las variantes léxicas de la entrada. Supone algunas conjeturas respecto de la "caja negra" del desarrollo software y operable online. (Ver detalles estructurales y funcionales en Figura 4).

La denominación de "modelo hipotético" se fundamenta en que, en tanto es una emulación analógica no sería esencialmente la única como proceso real que no esta transparentada en la bibliografía disponible de UMLS (solamente inferida metodológicamente a partir de tutoriales y publicaciones indirectas), y en cuanto a que sirve suficientemente para operatibilidad modélica de aproximación sistémica. Tal hipótesis será corroborable en la práctica de uso y en la eficacia con que en la práctica vaya siendo evaluada

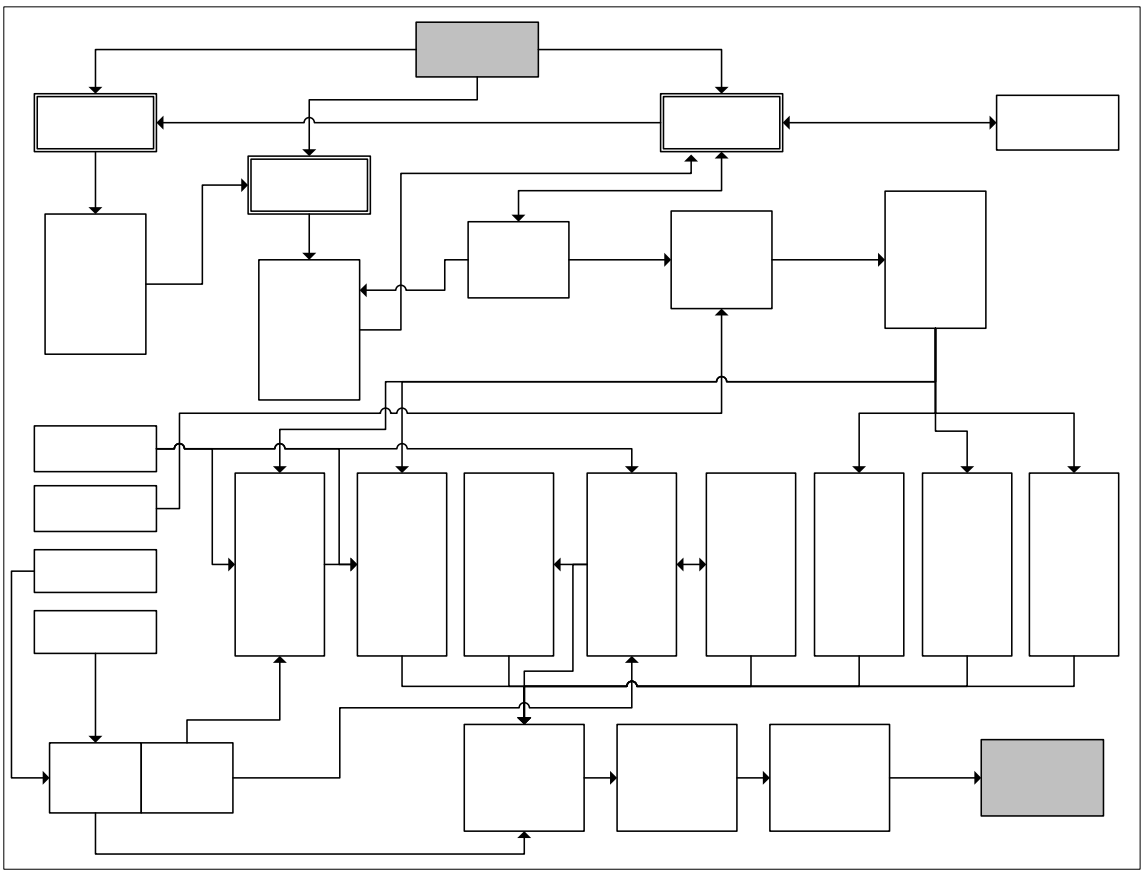

Figura 4: Modelo Hipotético # 1 (SPECIALIST lexicon UMLS KS, inglés, existente)

La mejor comprensión de los procesos intervinientes solo pueden ser entendidos desde un ejemplo comentado, sin que excedan los alcances de la presente comunicación. (Ver Figura 5).

Modelo hipotético # 2

Es un emulante del inglés al español, es por ello que su estructura se mantiene (no podría ser de otro modo) pero no su funcionalidad, la cual varía para el reconocimiento léxico del español. Es un modelo cuya complejidad está directamente relacionada con el idioma

español que tratará lexicalmente. Se lo define como el modelo *meta*, por ser la instancia de máxima a la que se aspira llegar luego de los refinamientos sucesivos de la prototipación beta. Los instrumentos lexicales que intervienen en este modelo son: **Norm** (Normalización); **WordInd** (Indice de palabras) y **LVG** (Generador de Variantes Léxicas). Luego de la intervención del instrumento lexical Norm, se genera un output reconvertido a mayúsculas, reemplaza puntuación por espacios, remueve fin de palabra, corrige ortografía, desinflexiona cada palabra y las ordena alfabéticamente y almacena el string normalizado en la tabla correspondiente.

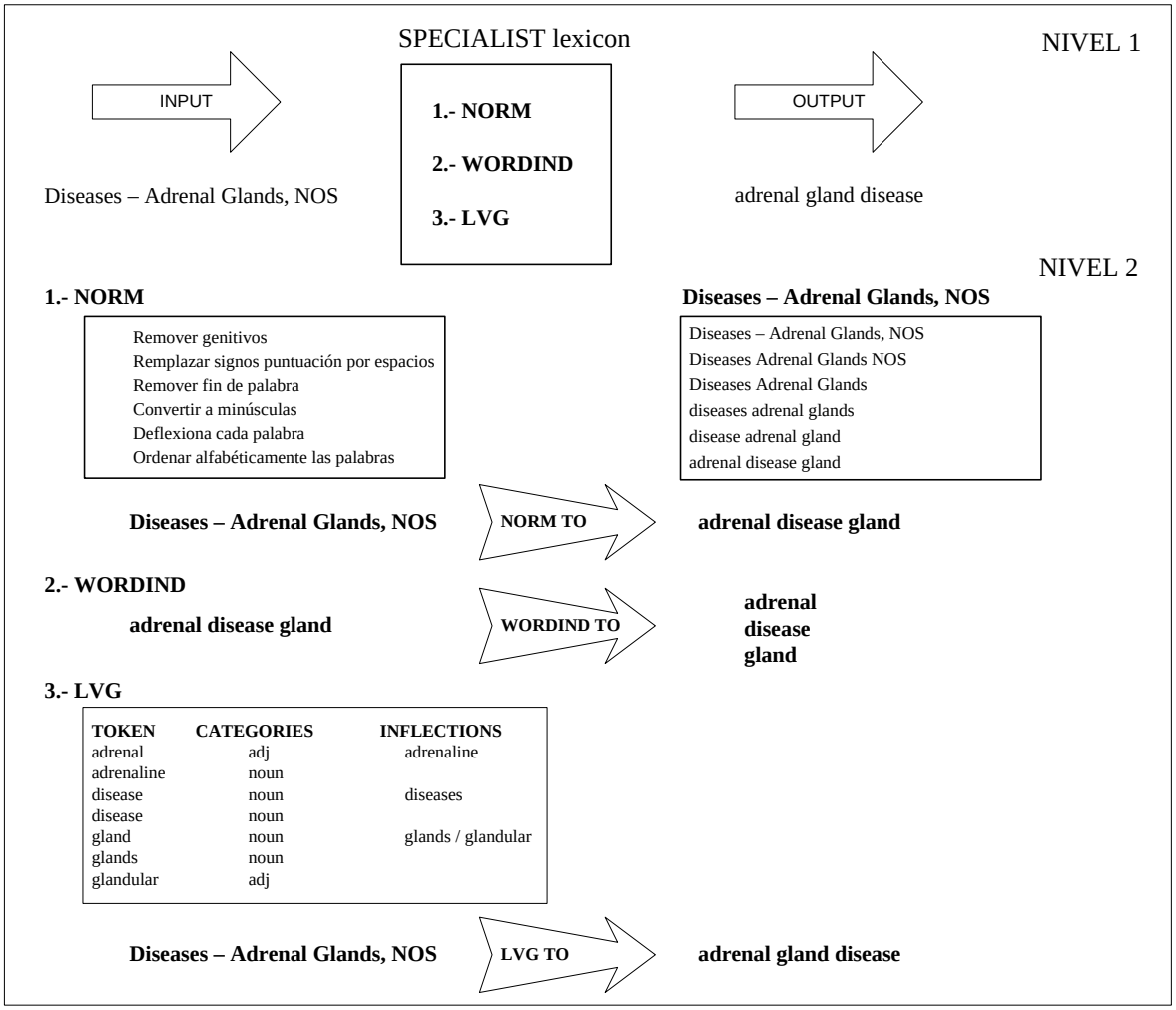

Figura 5: Modelo Hipotetico # 1 – Ejemplo de modelo lexical en inglés

El WordInd toma cada string como input y su función principal es la de tokenizar. Identificando cada palabra por los espacios en blanco, convierte a mayúsculas, pasa campos de entrada a campos de salida intactos, remueve espacios en blanco y signos de puntuación. El LVG es un instrumento lexical que genera las variantes léxicas para cada entrada. Establece para cada token la categoría sintáctica, obtiene las variantes inflexionadas y los patrones, determina el número de flujos que se generaron para determinar la variante de la entrada y por último obtiene el registro único, que es el registro para cada variante encontrada. Las tablas y sus atributos fueron redefinidos para el manejo

y almacenamiento de la información necesaria para los instrumentos lexicales a fin de encontrar las variantes. Es un modelo que reconoce todas las categorías sintácticas y determina la inflexión y derivación de las palabras que se encuentren en el lexicón. De no ser así debería generarlas algorítmicamente si existieren todas las reglas necesarias para realizarlas (Ver detalles representativos en Figura 6).

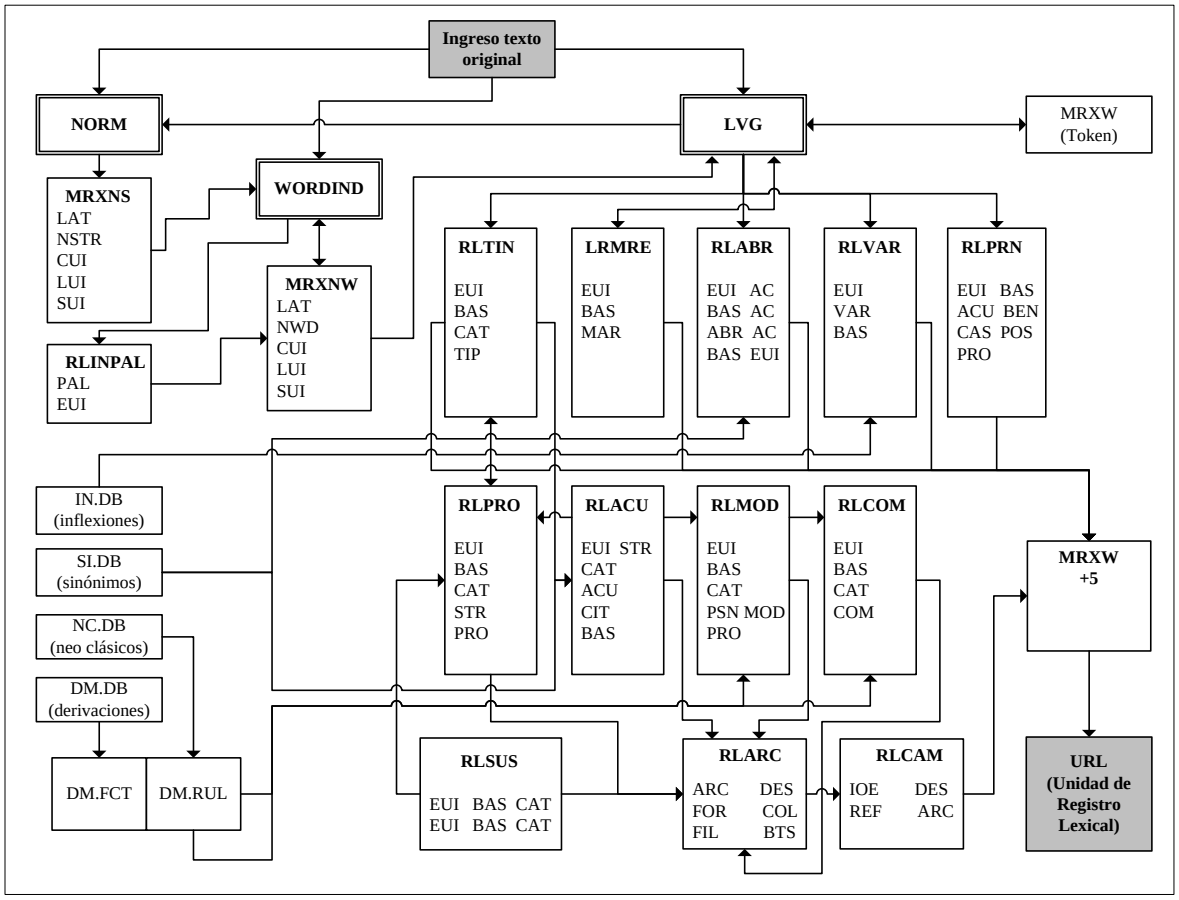

Figura 6: Modelo Hipotético # 2 (español, "complejo"/modelo meta)

Para este modelo fueron definidos los requerimientos lógicos del sistema sin imponer una implementación física. Las especificaciones fueron hechas para las tablas relacionales y auxiliares y para las bases de datos que intervienen en el.

Dentro de los lineamientos propuestos por AIPO a través de su metodología que proporciona la manera de proceder organizadamente para poder conseguir usabilidad en el diseño, considera a los prototipos como uno de los pilares básicos dentro de su modelo de proceso de ingeniería de la usabilidad y la accesibilidad centrado en el usuario. Al referirse a prototipos puede pensarse en documentos, diseños o sistemas que simulan partes del sistema final que se debe implementar. Una de las técnicas utilizadas para simular el comportamiento del sistema es el prototipado denominado "de lápiz y papel" (como exigencia mínima para el diseño) que sirven para garantizar que se cumplan los pasos necesarios para disponer de un producto altamente usable. (Ver ejemplo de corrida prototipada en Figura 7)

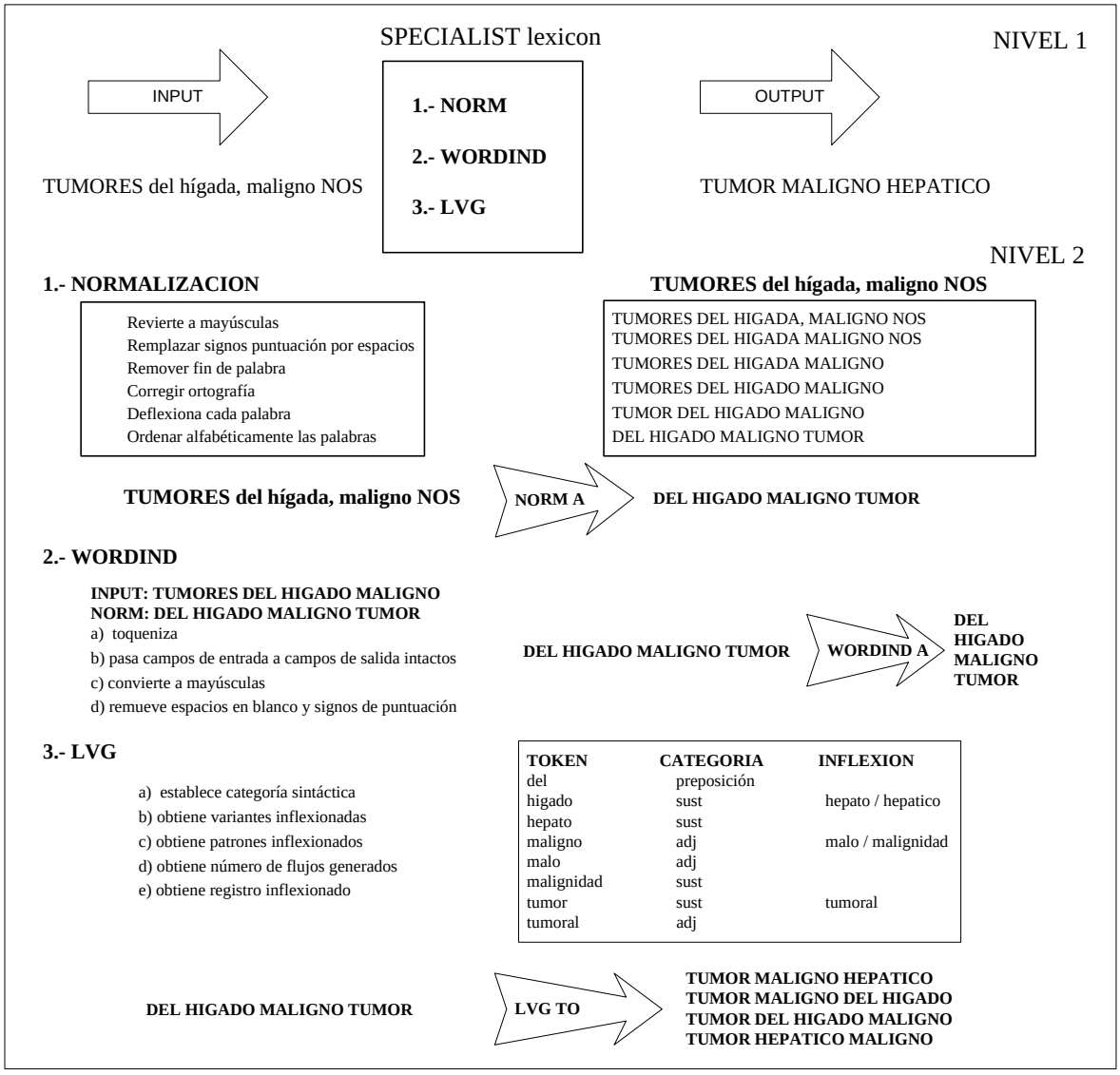

Figura 7: Modelo Hipotético # 2 – Ejemplo de proceso lexical en español

Modelo hipotético # 3

Debido a que la sistémica alienta la creatividad se ideó un modelo distinto al emulado. Es un modelo que maneja el español de manera más simple que el modelo hipotético 2, definido como meta. Diseñado para el reconocimiento por ahora de sintagmas nominales, debido a la complejidad diferencial del paradigma verbal y ante la evidencia de la empiria que no presenta esta categoría sintáctica en el dominio de la lista de problemas existente y disponible. Si bien se entiende que el diseño de la interfaz debe realizarse con cautela, porque ante mayores restricciones normalmente se generan mayores resistencias, este modelo propone un diseño de interfaz (tipo light orientativa al usuario) que permita determinar las condiciones de limitación/orientación/regulación del lenguaje coloquial, en la medida que la reducción de grados de libertad ayude a una reducción en los errores, lo cual no implica una restricción de libertad sino una corrección en el uso de lenguaje español. Es un modelo netamente operativo (pensando para la prototipacion recursiva) por

su simplicidad de diseño y es imprescindible para la ulterior prototipación y prueba en especial de las componentes lingüísticas, de los datos y de las tablas. El propósito principal de este modelo hipotético es el de constatar -lógica, secuencialidad, alcance, suficiencia- las *reglas gramaticales* definidas para determinar los sintagmas en sus categoría sintáctica y las inflexiones y variantes de sinonimia, en particular de la colección de los ingresos de la lista de problemas de Itálica. Se realizó un diagrama conjetural de la secuenciación de las **reglas** dentro de los desarrollos macro-implementación para la corrida del modelo hipotético 3 (Ver la conjetura diagramada en la Figura 8).

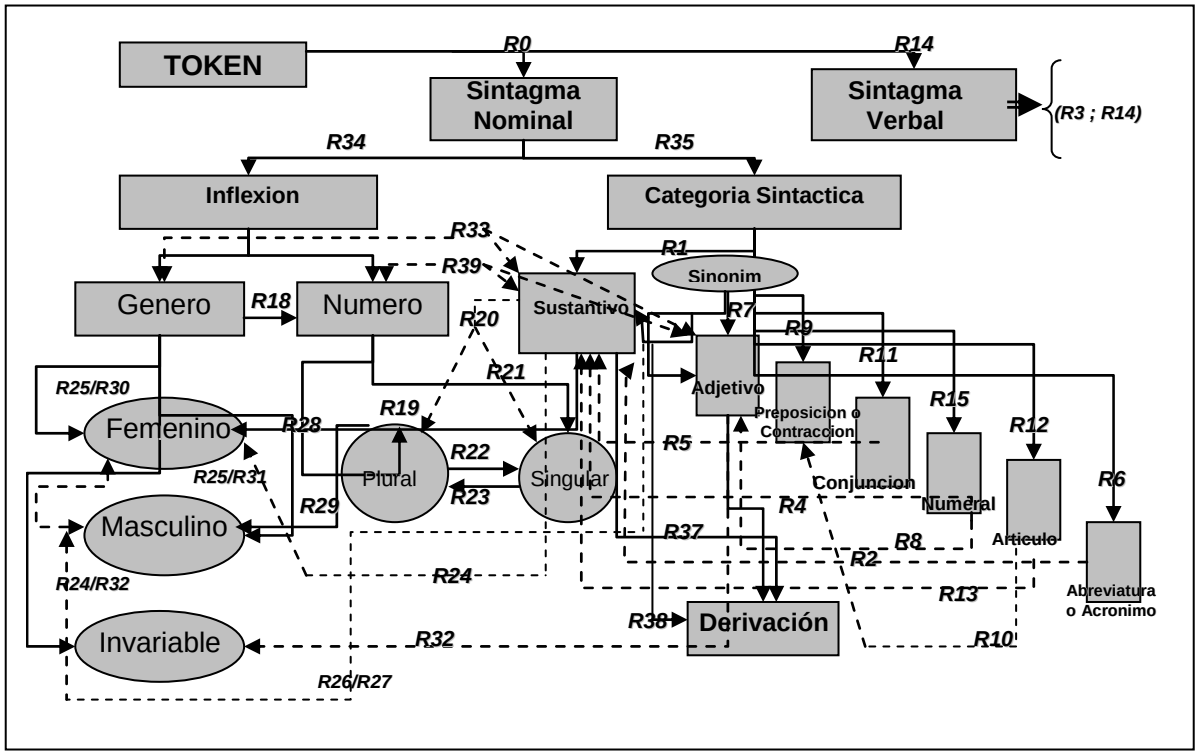

Figura 8: Reglas Gramaticales secuenciadas (conjetura solo para Modelo Hipotético 3)

Otra característica distintiva es que incorpora el Proto-Tesauro HIBA dentro del modelo evitando el tratamiento léxico de los problemas ingresados que ya se encuentran codificados en el Tesauro, también para su testeo. Es de destacar la necesidad de rescatar la empiria generada sobre la base de la experticia asistencial del HIBA que es tratada informáticamente por su HCE desde 1998 hasta el 2004, que fue generando una base de datos con textos codificados que engrosan el Proto-Tesauro con 5.963 conceptos, solo para la data 1998-2002. Esto permite, mediante estrategias de búsqueda y normalización de textos elementales, la comparación del texto ingresado por el usuario médico (que llega al millón y medio de ingresos) con la base de datos y de esta forma la asignación del código correspondiente. (Ver comparación de procesos en Figura 9)

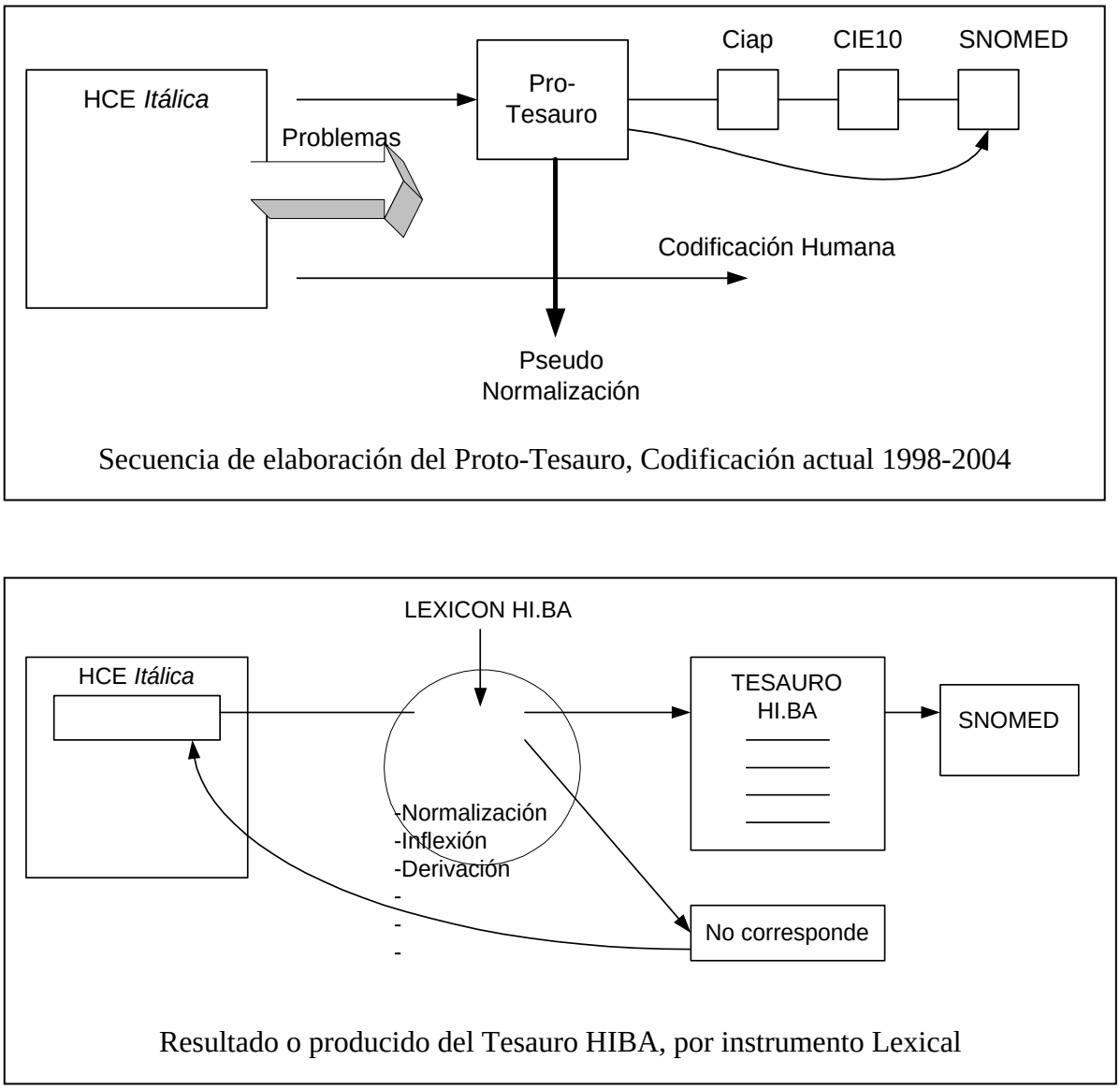

Figura 9: Comparación de procesos en Itálica

Síntesis de modelización sistémica

El tiempo dedicado al estudio del UMLS-KS [25, 29-31] para comprenderlo como marco de referencia en orden al procesamiento del lenguaje natural insumió la mayor dedicación, fue la base para el pre-diseño. Una vez comprendidas sus fuentes de conocimiento pudo profundizarse sistemicamente en la comprensión del sistema emulado, requerimiento principal para el desarrollo del proyecto. La emulación sistémica como metodología empleada permitió abrir y desagregar en distintos niveles de comprensión la “caja negra” del Process, mientras se profundiza en el conocimiento de Input y Output. Así se obtiene un primer modelo hipotético en inglés análogo al real, que es el existente y se deriva de este el emulante modelo hipotético 2 en español o castellano. Surge un modelo hipotético 3, operante, por la necesidad de probar las reglas gramaticales que mediante refinamientos recursivos de prototipación tenderá a convertirse a la larga en el modelo hipotético 2. De

esta manera se realiza una emulación de mínima, que funcione y una de máxima, buscando obtener resultados óptimos, con velocidad, precisión y mayor granularidad en un horizonte prospectivo. Mediante la prototipación de la hipótesis tres, se correrá próximamente el *prototipo beta*, como modelo operante que funciona en base a las reglas gramaticales cuyos resultados serán comparados con los resultados actuales que se procesan de manera semiautomática y con pasos manuales. (Ver el esquema síntesis de la modelización sistémica en la Figura 10).

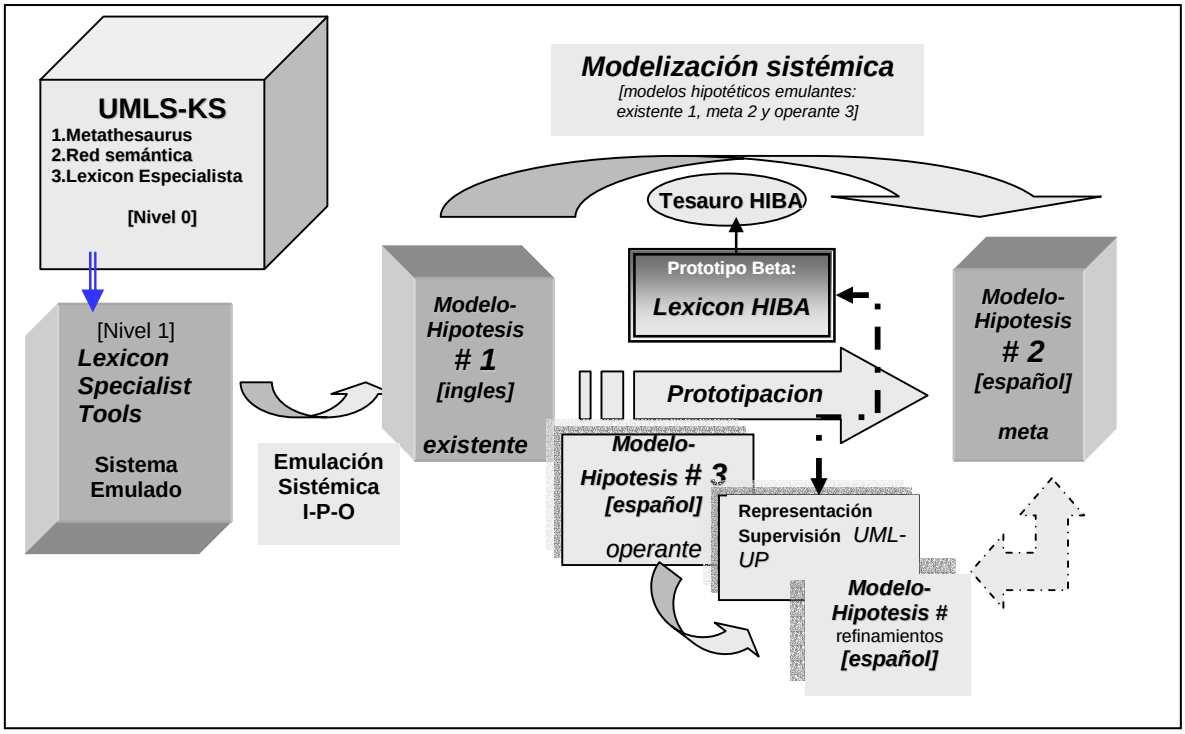

Figura 10: Síntesis de modelización sistémica

El manejo simultáneo de tres hipótesis permite seguir emulando, centrándonos en la relación existente entre los actores y los roles en el contexto de las tareas a realizar en lo que hace a la organización –que en su marcha evolutiva no se detiene en magnitud expansiva-, que garanticen el desarrollo de un producto con un grado mucho más alto en cuanto a la usabilidad y accesibilidad, además de la satisfactibilidad, de dicho producto. El marco contextual –que obviamente debe ser sistemicamente atendido y tratado- en si constituye un sistema mayor que va dotando de la necesaria coherencia y operacionalidad a los macro procesos, creativos y evolutivos, que se tienen en cuentra para el tratamiento informático de significaciones. Este macro-sistema se encuentra ya bosquejado y en mente.

Desarrollos lingüísticos

Los desarrollos más críticos se ubicaron en la formalización del tratamiento informático de los morfemas en español. Para esto se establecieron las necesarias y, al parecer por ahora, suficientes reglas ortográficas, morfológicas y sintácticas como pautas que se deben obedecer para el tratamiento algorítmico de los sintagmas introducidos en español (texto libre) por el médico en la lista de problemas de Itálica. En cuanto al tratamiento lingüístico

del español oficial, se trabajó en un primer momento con la determinación de clases de palabras [32, 33]. Para ello se procedió de dos maneras:

- Por un lado, se registraron las categorías sintácticas cerradas, es decir compuestas por un conjunto limitado de palabras, fáciles de reconocer (artículos, preposiciones, conjunciones)
- Por otro lado, se determinaron listados bastantes exhaustivos de inflexiones o derivaciones propias de cada clase de palabras (especialmente sustantivos y adjetivos). Se realizaron tablas con sufijos que permiten reconocer un sustantivo y un adjetivo.

Se llegó a determinar en esta instancia 40 reglas gramaticales, hasta el momento. Solo una mínima porción del corpus trabajado comprende términos individuales, que se encaran mediante el procedimiento de análisis morfológico. La gran mayoría de las entradas corresponde a frases conformadas por varias palabras, generalmente ***sintagmas nominales***, sin verbo. Por esta razón, se procedió a determinar reglas de las estructuras sintagmáticas en las que podían conformarse estas frases nominales, y luego las adjetivales y adverbiales, que a su vez pueden estar englobadas en las primeras. Las reglas desarrolladas permiten obtener información sobre las categorías gramaticales asociadas a las palabras que aparecen en la frase que se desea analizar, adjetivo, sustantivo, preposiciones, conjunciones, reconocer un artículo, reconocer un conector, acrónimos y abreviaciones y sinónimos. Permiten determinar la inflexión de las palabras, determinar el género y el número y establecer la derivación a partir de la obtención de la forma base o lema en forma directa.
En una fase posterior del trabajo, se abordan las entradas que contengan verbos, con sus consiguientes complementos (OD, OI, Circunstanciales, etc.)

Testeo y prueba del prototipo Beta

La demanda formulada en los requerimientos de contar con una máquina abstracta que responda a exigencias y requisitos evolutivos, orientó hacia la prototipación [34]. Los refinamientos sucesivos pueden garantizar ajustes sobre la base de diseños flexibles. Que funcione para una gramática automática y que ese funcionamiento disminuya la variabilidad, y que sea un sistema evolutivo para niveles de mayor complejidad, es una mera aspiración si no se asumen recaudos desde el mismo inicio del proyecto. La cuestión metodológica fue y es central. Mediante la prototipación de la hipótesis tres, se correrá el *prototipo beta*, como modelo operante que funciona sobre la base de las reglas gramaticales cuyos resultados serán comparados contra los resultados actuales que se procesan de manera semiautomática y con pasos manuales que generaron el Proto-Tesauro HIBA, con sus resultados, alcances y parámetros. Este testeo supondrá la validación y verificación imprescindible a la modelización. La puesta a prueba del prototipo (prueba beta) se encuentra en pleno desarrollo. Los parámetros de testeo (validación y verificación) son realizados sobre la base de los rendimientos del procesamiento semiautomático en paralelo para la información médica generada durante Enero-Junio 2004, siendo aproximadamente 120.000 problemas. Las exigencias se fijan en superar los rangos del 80 % de precisión en el manejo del Proto-Tesauro HIBA (codificación CIAP2 y CIE 10 y SNOMED CT) en el lapso 1998/2003. La instancia de prueba brindará los resultados comunicables estadísticamente en breve.

Conclusión

La necesidad de dotar a Itálica con más sistemas clínicos de soporte para la toma de decisiones por un lado e integrar las diferentes terminologías clínicas utilizadas en los sistemas de información del HIBA llevó a la creación de un servidor de terminología. Uno de los componentes del mismo esta constituido por una pieza de software que se encarga del tratamiento lexical de los ingresos de texto narrativo en la interfaz del usuario de las aplicaciones. Para lograr dicha funcionalidad se utilizó la *"emulación sistémica"* del SPECILIST lexicon del UMLS en ingles, para adecuarlo al español. Se optó por la prototipación, la cual implicó un esfuerzo adicional en el pre-diseño del sistema ya que insumió mucho tiempo la búsqueda de antecedentes y exploración de bibliografía y de casos semejantes. Una dificultad encontrada y que no se pudo superar, dejándola para una etapa posterior, fue el desarrollo de reglas gramaticales adecuadas que permitan el reconocimiento de sintagmas verbales debido a la complejidad del paradigma verbal del español. Situación similar se documenta en trabajos similares tanto para el español [35], como para el francés [36] y el alemán [37], que también se vieron limitados a sintagmas nominales en sus primeras etapas. Esperamos que el esfuerzo de crear estos instrumentos lexicales para el manejo automático del español natural brinde sus máximos beneficios cuando se pasen a los desarrollos semánticos en un futuro próximo.

Referencias

[1] Rector, A.L., *Clinical terminology: why is it so hard?* Methods Inf Med, 1999. 38(4-5): p. 239-52.

[2] Sittig, D.F., *Grand challenges in medical informatics?* J Am Med Inform Assoc, 1994. 1(5): p. 412-3.

[3] Spackman, K.A., Campbell, K.E.Cote, R.A., *SNOMED RT: a reference terminology for health care.* Proc AMIA Annu Fall Symp, 1997: p. 640-4.

[4] Cornet, R.Prins, A.K., *An architecture for standardized terminology services by wrapping and integration of existing applications.* Proc AMIA Symp, 2003: p. 180-4.

[5] OMG, Lexicon Query Service, version 1.0. Acceced: 10 Jun 2004 [http://www.omg.org/docs/formal/00-06-31.pdf].

[6] Hogarth, M.A., Gertz, M.Gorin, F.A., *Terminology Query Language: a server interface for concept-oriented terminology systems.* Proc AMIA Symp, 2000: p. 349-53.

[7] HL7, Common Terminology Services, Version 0.8. Acceced: 10 Jun 2004 [http://www.hl7.org/library/committees/vocab/ctsspecv08.zip].

[8] Chute, C.G., Elkin, P.L., Sherertz, D.D.Tuttle, M.S., *Desiderata for a clinical terminology server.* Proc AMIA Symp, 1999: p. 42-6.

[9] Rose, J.S., Fisch, B.J., Hogan, W.R., Levy, B., Marshal, P., Thomas, D.R.Kirkley, D., *Common medical terminology comes of age, Part One: Standard language improves healthcare quality.* J Healthc Inf Manag, 2001. 15(3): p. 307-18.

[10] McCray, A.T., Aronson, A.R., Browne, A.C., Rindflesch, T.C., Razi, A.Srinivasan, S., *UMLS knowledge for biomedical language processing.* Bull Med Libr Assoc, 1993. 81(2): p. 184-94.

[11] McCray, A.T., *The nature of lexical knowledge.* Methods Inf Med, 1998. 37(4-5): p. 353-60.

[12] Lindberg, D.A., Humphreys, B.L.McCray, A.T., *The Unified Medical Language System.* Methods Inf Med, 1993. 32(4): p. 281-91.

[13] Gomez, A., Bernaldo de Quiros, F.G., Garfi, L., Luna, D., Sarandria, G., Figar, A., Martinez, M., Campos, F.D., K. *Implementación de un sistema de mensajeria electrónica -HL7- para la integración de un sistema multiplataforma*. in *4to Simposio de Informática en Salud - 30 JAIIO*. 2001. Buenos Aires, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO).

[14] Luna, D., Otero, P., Gomez, A., Martinez, M., García Martí, S., Schpilberg, M., Lopez Osornio, A.Bernaldo de Quiros, F.G. *Implementación de una Historia Clínica Electrónica Ambulatoria: "Proyecto ITALICA"*. in *6to Simposio de Informática en Salud - 32 JAIIO*. 2003. Buenos Aires, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO).

[15] Gonzalez Bernaldo de Quiros, F., Soriano, E., Luna, D., Gomez, A., Martinez, M., Schpilberg, M.Lopez Osornio, A. *Desarrollo e implementación de una Historia Clínica Electrónica de Internación en un Hospital de alta complejidad*. in *6to Simposio de Informática en Salud - 32 JAIIO*. 2003. Buenos Aires, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO).

[16] Luna, D., Bernaldo de Quiros, F.G., Garfi, L., Soriano, E.O´Flaherty, M., *Reliability of secondary central coding of medical problems in primary care by non medical coders, using the International Classification of Primary Care (ICPC).* Medinfo, 2001. 10(Pt 2): p. 300.

[17] Lopez Osornio, A., Luna, D.Bernaldo de Quiros, F.G. *Creación de un sistema para la codificación automática de una lista de problemas*. in *5to Simposio de Informática en Salud - 31 JAIIO*. 2002. Santa Fe, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO).

[18] Lopez Osornio, A., Montenegro, S., García Martí, S., Toselli, L., Otero, C., Tavasci, I., Luna, D., Gomez, A.Gonzalez Bernaldo de Quiros, F. *Codificación múltiple de una lista de problemas utilizando la CIAP-2, CIE-10 y SNOMED CT*. in *3er Virtual Congress of Medical Informatics - Informedica*. 2004.

[19] Pollán, J., Arbelbide, J., Pedernera, F., Borbolla, D., Achilli, F., Victoria, V., Gonzalez Bernaldo de Quiros, L., Gomez, A., Luna, D.Gonzalez Bernaldo de Quiros, F. *Medición de la calidad de los registros codificados en una historia clínica electrónica de internación*. in *3er Virtual Congress of Medical Informatics - Informedica*. 2004.

[20] Otero, P., Bernaldo de Quiros, F.G., Luna, D., Garfi, L., Gomez, A., Martinez, M.Staccia, G. *Desarrollo e implementación de un sistema estructurado de solicitud de exámenes complementarios desde una Historia Clínica Electrónica Ambulatoria*. in *4to Simposio de Informática en Salud - 30 JAIIO*. 2001. Buenos Aires, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO).

[21] Luna, D., Bernaldo de Quiros, F.G., Garfi, L., Morchón, A., Gomez, A., Martinez, M.Staccia, G. *Unidad asistencial: Creación de un nueva clasificación para la implementación de un sistema de prescripción electrónica*. in *4to Simposio de Informática en Salud - 30 JAIIO*. 2001. Buenos Aires, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO).

[22] Morchón, A., Pedernera, F., Otero, P., Costa, G., Lopez Noguerol, M., Martinez, M., Gomez, A., Gassino, F., Lopez Osornio, A., Luna, D.Gonzalez Bernaldo de Quiros, F. *Desarrollo de un vocabulario para dispositivos médicos*. in *7to Simposio de Informática en Salud - 33 JAIIO*. 2004. Córdoba, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO) - Enviado a referato.

[23] Schpilberg, M., Bernaldo de Quiros, F.G., Luna, D., Gomez, A., Martinez, M.Cifarelli, G. *Creación de un sistema para la detección de interacciones farmacológicas en una Historia Clínica Electrónica*. in *4to Simposio de Informática en Salud - 30 JAIIO*. 2001. Buenos Aires, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO).

[24] Luna, D., Hares, D., Schpilberg, M., Hernandez, G., Soriano, E., Martinez, M., Gomez, A., Cifarelli, G.Bernaldo de Quiros, F.G. *Validación de la base de conocimiento de un sistema notificador de interacciones farmacológicas*. in *5to Simposio de Informática en Salud - 31 JAIIO*. 2002. Santa Fe, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO).

[25] National Library of Medicine, *The Specialist Lexicon and Lexicon Programs*, in *UMLS Knowledge Sources*. 2003: July Release 2003AB.

[26] SIGCHI (Group : U.S.). Curriculum Development Group., *ACM SIGCHI curricula for human-computer interaction*. 1996, New York: Association for Computing Machinery. iv, 162 p.

[27] Lorés, J., ed. *La Interacción Persona-Ordenador*. 1ra ed. 2001, AIPO, Asociación Interacción Persona Ordenador.

[28] Fowler, M.Scott, K., *UML distilled : a brief guide to the standard object modeling language*. 2nd ed. 2000, Reading, Mass.: Addison Wesley. xxi, 185 p.

[29] Bodenreider, O., Hole, W.T., Humphreys, B.L., Roth, L.A., Srinivasan, S., *Customizing the UMLS Metathesaurus for your applications*, Tutorial AMIA Fall Symposium, Nov 9 2002, San Antonio, Texas. Acceced: 10 Jun 2004 [http://etbsun2.nlm.nih.gov:8000/pres-ob/021109-AMIA-tutorial/T13-color.pdf].

[30] Broune, A., Divita, G., Lu, C., National Library of Medicine, *Lexicon Tools for UMLS Developers*, Tutorial AMIA Fall Symposium, Nov 9 2002, San Antonio, Texas. Acceced: 10 Jun 2004 [http://umlslex.nlm.nih.gov/lvg/current/docs/userDoc/tutorial/index.html].

[31] Lenci, A., *Ontologies and the Lexicon.* Università di Pisa - Department of linguistics & Istituto di Linguistica Computazionale. Tutorial in Course on Language & Ontology Copenhagen Business School. Acceced: 10 Jun 2004 [http://www.humaniora.sdu.dk/ifki/ontoquery/presentations/Ontologies-Lexicon-1.ppt].

[32] Bosque, I. Demonte, V., *Gramática descriptiva de la lengua española*. 1ra ed. Colección Nebrija y Bello. 1999, Madrid: Espasa Calpe.

[33] Real Academia Española., *Diccionario de la lengua española*. 22. ed. 2001, Madrid: Editorial Espasa Calpe. 2 v. (lvii, 2368 p.).

[34] Vonk, R., *Prototyping : the effective use of CASE technology*. 1990, New York: Prentice Hall International. xiv, 157 p.

[35] Berra, M.C., Creteur, M., Strobietto, R., March, A.Reynoso, G. *Desarrollo de un Lexicón y de herramientas léxicas para la búsqueda y recuperación de información médica en español*. in *4to Simposio de Informática en Salud - 30 JAIIO*. 2001. Buenos Aires, Argentina: Sociedad Argentina de Informática e Investigación Operativa (SADIO).

[36] Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch, P., Le Duff, F., Thirion, B.Darmoni, S., *UMLF: A Unified Medical Lexicon for French.* Proc AMIA Symp, 2003: p. 1062.

[37] Weske-Heck, G., Zaiss, A., Zabel, M., Schulz, S., Giere, W., Schopen, M.Klar, R., *The German specialist lexicon.* Proc AMIA Symp, 2002: p. 884-8.

Datos de Contacto:

Ing.: Maria Mercedes Clusella Cornejo.
Colegio Mayor Universitario Santiago del Estero
Grupo de ingeniería de Software.
Avda. Belgrano (s) 1915 UF 202, (G4200ABG)
Ciudad de Santiago del Estero, Argentina
E-mail: ColegioMayorUniversitarioSdE@Argentina.com