STILUS: Sistema de Revisión Lingüística de Textos en Castellano

Julio Villena, Beatriz González, José Carlos González, María Muriel

DAEDALUS, S.A. – Paseo de las Delicias 31, 3º
28045 Madrid (España)
{jvillena, bgonzalez, jgonzalez, mmuriel}@daedalus.es

Resumen. Este documento presenta la descripción general y la arquitectura de STILUS, un sistema modular de revisión lingüística de textos en castellano, diseñado para alcanzar un elevado grado de precisión y cobertura de fenómenos lingüísticos en este idioma, incluyendo errores gramaticales, ortográficos y de estilo. STILUS comenzó como un proyecto de I+D financiado por la iniciativa ATYCA del Ministerio de Industria en 1999 (STILUS: Servicios Telemáticos de Ingeniería Lingüística) y ha evolucionado hasta convertirse en un sistema comercial.

1 Introducción

La revisión lingüística de textos es la aplicación de la ingeniería lingüística que tiene como objetivo la detección de los errores lingüísticos cometidos en un texto escrito en lenguaje natural. Hoy en día los resultados de este campo se plasman principalmente en el desarrollo de sistemas de revisión automática de textos, integrados en herramientas de edición de documentos, aunque va a cobrar mucha más importancia en el futuro con la evolución de los interfaces avanzados persona-máquina, como complemento de los sistemas de reconocimiento automático de voz.

Aquí se presenta la descripción general y la arquitectura de STILUS, un sistema de revisión lingüística orientado a textos en castellano, diseñado para obtener un elevado grado de precisión y cobertura de fenómenos lingüísticos en este idioma. STILUS comenzó como un proyecto de I+D financiado por la iniciativa ATYCA del Ministerio de Industria en 1999 (STILUS: Servicios Telemáticos de Ingeniería Lingüística), aunque con el tiempo ha evolucionado hasta convertirse en un sistema comercial (www.revisado.com).

Básicamente se distinguen cuatro tipos de errores: *gramaticales, ortográficos, semánticos* y *de estilo*. El sistema dispone de módulos expresamente dedicados a cada uno de ellos. En el caso de los errores semánticos, la detección superficial que realiza el sistema se basa en criterios morfosintácticos o en la aparición de determinadas estructuras, en cierta forma asimilando la detección de errores semánticos a los otros tipos de errores [5].

2 Arquitectura del Sistema

El sistema tiene una arquitectura modular que se compone de diferentes bloques encadenados, que van actuando sobre la salida del bloque anterior.

Los módulos previos son la *detección del idioma* y el *filtrado* que convierte los documentos a un formato de representación interno con el que trabajan el resto de los módulos, llamado FTA (Formato Tabular Ampliado). Seguidamente, el módulo de segmentación divide el texto en unidades, a las que el módulo de análisis morfosintáctico asigna su etiqueta correspondiente, utilizando el etiquetario definido y representando los resultados en el mismo formato FTA. Sobre esta salida actúa el módulo de *revisión lingüística*, que se descompone en tres submódulos principales: la revisión ortográfica, la revisión gramatical y la revisión de estilo.

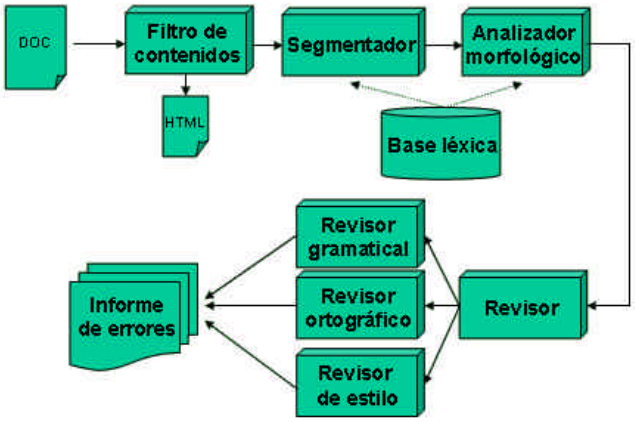

Fig. 1. Arquitectura del sistema

3 Identificación del Idioma y Filtrado

La utilidad de la identificación del idioma es impedir la revisión de textos escritos en idiomas distintos del castellano. El análisis se basa en la detección de palabras más frecuentes de los idiomas reconocidos.

El objetivo fundamental del filtrado es tratar la mayor cantidad posible de formatos electrónicos. El proceso de filtrado realiza una primera conversión de los formatos electrónicos al formato HTML mediante el empleo de filtros externos (más de 250 formatos diferentes), y luego convirtiéndolo al formato interno FTA, extrayendo información de tipografía (negrita, cursiva, subrayado, título), la posición dentro del documento original, saltos de línea y de párrafo y realizando las conversiones necesarias entre juegos de caracteres.

4 Segmentación y Etiquetado

La segmentación es el proceso por el cual un texto se divide en unidades de varios tipos para ser etiquetados posteriormente de acuerdo con su categoría o función morfosintáctica.

Además, en esta fase se detecta la función de las diferentes unidades: guiones (separar palabras, fin de línea, etc.), símbolos, signos de puntuación, letras y números mezclados, etc. Gracias a los recursos de los que dispone el sistema, prácticamente todas las palabras reciben una categorización específica adecuada. Una dificultad importante dentro de la fase de etiquetado es la existencia de palabras desconocidas, que se solventa mediante un proceso constante de enriquecimiento de diccionarios.

4.1 Base Léxica

La base léxica contiene el total de las palabras reconocidas como válidas por el analizador morfológico. A partir de ella, se generan diccionarios objeto, compilados y optimizados por herramientas específicas para su acceso y consulta desde aplicaciones cualesquiera.

Las dimensiones de la base léxica son las siguientes:

Tabla 1. Dimensiones de la base léxica

Categoría gramatical	Entradas
Sustantivos	37.270
Adjetivos	15.730
Verbos	8.325
Categorías cerradas (adv., prep., conj., art., etc.)	1.707
Formas verbales lexicalizadas	226
Nombres propios	24.969
TOTAL	**88.227**

A partir de cada entrada se generan, por procedimientos automáticos y de acuerdo con modelos preestablecidos, todas las raíces posibles (por ejemplo, para el verbo *haber*, se generan las raíces *hab-*, *hub-* y *hay-*), que pueden ser combinadas correctamente con algunos de los 622 morfemas flexivos. La base léxica incorpora expresiones multipalabra que constituyen una unidad desde el punto de vista sintáctico. Por ejemplo: “a costa de”, “Juan Carlos I”, “con respecto a”, etc.

Hay tres tipos de entradas léxicas:

- *Lemas:* Son las formas paradigmáticas que resumen el conjunto de formas flexivas que pueden derivarse mediante un proceso morfológico. Para facilitar esta tarea se definen unos modelos flexivos y unas reglas de alomorfía aplicables a cada uno de ellos. Estas reglas permiten la generación de las posibles variantes alomórficas (siempre de las raíces de las formas flexivas), cada una con los rasgos pertinentes

asociados. El formalismo incluye herramientas que permiten obtener automáticamente los alomorfos correspondientes para un lema dado.

- *Morfemas:* Sufijos flexivos y sus alomorfos, que se concatenan con los alomorfos asociados a los lemas de la base léxica durante el proceso morfológico. Generalmente llevan asociados rasgos de tipo sintáctico que hereda la palabra formada con ellos.
- *Formas lexicalizadas:* Son formas que no se pueden obtener directamente del proceso morfológico, normalmente por tratarse de formas flexivas muy irregulares o formas nominales de género o número inherente.

La base léxica da cuenta de más de 1.700.000 formas distintas de palabras en castellano, donde se incluyen todas las posibilidades de afijación de pronombres enclíticos verbales. La depurada caracterización morfológica de todas las entradas del diccionario hace que no se sobrerreconozcan entradas (no se den por buenas combinaciones incorrectas de raíces y morfemas).

La base léxica se complementa por otros diccionarios. El diccionario de frecuencias tiene como entradas las palabras más frecuentes del español y su correspondiente frecuencia de aparición. Los diccionarios temáticos son extensiones de la base léxica y contienen términos empleados en determinadas áreas de conocimiento: economía, música, tauromaquia, vocabulario jurídico, científico, etc.

4.2 Etiquetario

El etiquetario empleado es posicional, de tal forma que cada posición está reservada para un determinado rasgo pertinente en la categoría. En dicha posición sólo pueden aparecer los valores asociados a ese atributo.

Para cada clase de palabra se especifican mediante pares atributo/valor las características morfosintácticas de toda la clase. La aplicabilidad de todos los pares atributo/valor que caracterizan a una determinada categoría no es uniforme, por lo que hay atributos que sólo se aplican a ciertas clases. La no aplicabilidad de un rasgo a una determinada unidad léxica queda reflejada mediante el valor NIL.

En las descripciones morfosintácticas se dan cuenta de los cuatro tipos de información a que se refieren los atributos: información *léxica* (categoría, polaridad, femenino con 'a' tónica…), *morfológica* (género, número, persona, tiempo…), *funcional* (pronominal o no pronominal) y *textual* (si se trata de una multipalabra, una abreviatura, etc.).

Por ejemplo, la etiqueta para la palabra *"calendario"* es [NC--MS-N], donde:

N: categoría (sustantivo).

C: tipo (nombre común).

M: género (masculino).

S: número (singular).

N (final): indica que la unidad léxica está compuesta por una sola palabra que no presenta ninguna característica especial, como puntos o guiones.

Los guiones son rasgos que no se aplican a esta palabra (NIL).

4.3 Analizador Morfológico

El analizador morfológico se basa en un modelo de procesamiento morfológico que surge a partir del modelo de representación de la información lingüística definido en la plataforma ARIES [4]. Los análisis que se obtienen con este analizador consisten en la/s categoría morfológica/s que puede adoptar la palabra, representadas por una etiqueta.

En el modelo de ARIES, las palabras están compuestas de uno o más formantes. En general, una palabra podría componerse de un formante ("farol"), de dos ("niñ-o"), o de más ("niñ-it-o"). Aquí llamaremos raíz al primer formante (o único, en su caso), y terminación al segundo.

Cada formante lleva cierta información morfológica de rasgos, que será empleada para generar el/los análisis morfológico/s de la palabra completa (puede salir más de uno). Esta información consiste (en general) en un haz de rasgos morfosintácticos, como el género, el número, la persona, el lema de la palabra, etc.

Además, cada formante lleva cierta información morfológica de concatenación que indica con qué otros formantes se puede concatenar, pues no todos se pueden unir con todos. Por ejemplo, el formante "o" con información de rasgos de "género masculino, número singular" es distinto al formante "o" con información de rasgos de "1ª persona singular, presente indicativo", y por eso deben tener diferente información de concatenación: el primero tendrá concatenación con raíces nominales, y el segundo con raíces verbales que sean además de 1ª conjugación.

5 Revisor Ortográfico

El revisor ortográfico lleva a cabo la corrección de palabras erróneas en tres etapas.

5.1 Generación de Alternativas

El primer paso consiste en la generación de un conjunto de alternativas a la palabra errónea, mediante la aplicación de las cuatro operaciones básicas de edición (OBE):

- *Adición.* Inserción de letras en la palabra. Para una mayor eficiencia, se prueba únicamente con las letras posibles en ese contexto lingüístico y no con todas las letras en todas las posiciones.
- *Eliminación.* Para los errores dobles, se elimina una letra y se realiza un nuevo proceso de eliminación sobre la palabra generada.
- *Sustitución.* Técnica compuesta por una eliminación y una adición.
- *Transposición.* Intercambio entre letras adyacentes.

5.2 Ponderación de Alternativas

Para ordenar la lista de las alternativas, se emplean dos técnicas de ponderación combinadas con un coeficiente previo obtenido durante la etapa de generación: la ponderación por *frecuencias sintácticas* y la ponderación por *frecuencias de aparición*.

La primera técnica se basa en la frecuencia de aparición de las categorías sintácticas de las palabras incluidas en la lista de alternativas, teniendo en cuenta su inmediato contexto anterior y posterior.

A modo de ejemplo, se supone la siguiente oración:

*El equipo no pudo abrir la **defenda** férrea del contrario hasta el minuto 58.*

Teniendo en cuenta únicamente el conjunto de alternativas simples, la errónea '*defenda*' puede corregirse mediante las palabras siguientes:

defienda, defensa, defendía, dependa, defendí

El tratamiento de palabras aisladas trataría de corregir '*defenda*' ignorando el resto de las palabras de la oración, lo que le obligaría a elegir entre las alternativas anteriores basándose en otros criterios. La ponderación por frecuencias sintácticas toma en cuenta otras palabras del entorno de la palabra a corregir, tomando al menos las palabras anteriores y posteriores a la palabra errónea y, de una en una, las alternativas de la lista, para generar n-gramas. Por ejemplo, con trigramas:

la defienda férrea → *(artículo, verbo, adjetivo)*
la defensa férrea → *(artículo, nombre, adjetivo)*

De todos ellos, en la tabla de frecuencias sintácticas de trigramas, la segunda combinación *(artículo, nombre, adjetivo)* será más probable, así que se pondera favorablemente la alternativa '*defensa*' en detrimento del resto.

La segunda técnica, la ponderación por *frecuencias de aparición*, se basa en la frecuencia relativa de aparición en un corpus de las palabras incluidas en la lista de alternativas a ordenar. Cuando se ha detectado el error y generado las posibles alternativas, se comprueba la frecuencia de aparición de un corpus, y se ponderan las puntuaciones en función de dicha frecuencia.

5.3 Ordenación de Alternativas

La etapa final consiste en un ordenamiento de mayor a menor de las alternativas combinando los coeficientes obtenidos en las etapas anteriores.

6 Revisor Gramatical

El revisor gramatical toma del segmentador una pila con las unidades de segmentación del texto, que contiene además el análisis morfológico que han recibido las distintas unidades.

Seguidamente, se ejecutan sobre esta pila las reglas gramaticales, comprobando las restricciones que en ellas se definen, y una vez detectado el incumplimiento de una restricción, se inserta el mensaje de error en la pila de mensajes de error.

La definición de las reglas y el algoritmo de ejecución se derivan del expresado en el sistema CON-TEXT [11]. Las reglas gramaticales se componen por estructuras, restricciones y mensajes de error, y hacen uso de las definiciones variables y booleanas.

- *Definición de estructuras de la regla* (numeradas de 0 a n-1). Para activar una regla, se han de cumplir todas sus estructuras obligatorias y de inicio, de forma que cada unidad de segmentación de la oración coincida con cada una de las estructuras. En este paso se genera la nueva pila de unidades que se llamará desde ahora pila de estructuras. Si no se cumpliesen las estructuras obligatorias o iniciales, la regla dejaría de operar y se pasa a comprobar la siguiente regla.
- *Restricciones.* Comparaciones entre rasgos de las unidades que componen cada estructura, a modo de ecuación de rasgos. La restricción se cumple si se verifica la ecuación.
- *Mensajes de error* (opcional). Contiene el mensaje de error asociado a la regla.

6.1 Tipología de Errores Gramaticales

Se considera error gramatical toda transgresión de los principios y restricciones que rigen los cuatro niveles de análisis lingüístico: nivel de la palabra, nivel morfológico, nivel sintáctico, nivel semántico.

La elaboración de la tipología de errores se ha efectuado de manera empírica. Por un lado, se han extraído errores de contextos lingüísticos reales, acudiendo para ello a corpus, y, por otro, se ha reforzado con datos procedentes de distintos manuales de estilo o diccionarios normativos sobre el uso de la lengua. Las reglas gramaticales se crean a partir de esta tipología.

Los errores sintácticos se han dividido en dos grandes grupos:

- *Errores de concordancia*: contravienen restricciones de las categorías gramaticales de género y de número (intrasintagmáticos o intersintagmáticos).
- *Errores de secuencias*: errores que violan restricciones de secuencialización de las categorías léxicas.

Alrededor de estos dos grupos de errores se articula una gran variedad de errores sintácticos.

Los errores de secuencias transgreden restricciones de secuencialización de las categorías léxicas. Estas secuencias erróneas pueden estar constituidas por secuencias ilegales de categorías (léxicas), secuencias ilegales de categorías léxicas y restriccio-

nes gramaticales (categorías gramaticales), o simplemente de secuencias ilegales de palabras.

Esta clase de error se compone de muy variados subtipos, siendo el factor común el hecho de que producen secuencias ilegales: homofonía, grupos verbales continuos, preposiciones regidas por núcleos predicativos, dequeísmo y queísmo, sustitución de preposiciones, secuencias ilegales con amalgamas y cambio de forma en las conjunciones coordinantes.

Por último, se revisan los errores léxicos. En la lengua escrita se producen confusiones en el paradigma flexivo morfológico al que pertenecen determinadas unidades léxicas, ya que las confusiones están muy extendidas y aceptadas en la lengua hablada [10].

Estas confusiones pueden afectar a la formación de las formas flexivas irregulares de algunos verbos (*andaste* por *anduviste*, *dormió* por *durmió*...), o de algunos sustantivos (*déficits* como plural de *déficit*, *clubs* por *clubes*...).

7 Revisor de Estilo

Se consideran errores de estilo aquellas formas que dificultan la comprensión de un texto cuyo contenido y finalidad se encuadra dentro de un determinado uso de la lengua escrita. Para el sistema, el control del estilo significa la comprobación de la consistencia de un determinado texto a distintos niveles:

- Consistencia en el uso del sublenguaje.
- Consistencia en el uso de convenciones de contenido (abreviaturas, siglas, acrónimos, fechas, etc.).
- Errores de puntuación.

Para ello es necesario, por un lado, el procesamiento del formato del texto que se vaya a revisar, pero, por otro, el procesamiento del contenido del texto para averiguar cuáles son las reglas pertinentes que han de aplicarse al texto, lo que solamente puede hacerse si se realiza un análisis semántico del contenido del texto.

8 Evaluación

8.1 Revisión Ortográfica

Para evaluar la precisión y cobertura de los resultados de la revisión ortográfica, se presenta una evaluación comparativa con el corrector ortográfico de Microsoft Word XP. Se ha revisado ortográficamente con ambas herramientas un texto de evaluación (la versión digital de El País del 17 de Mayo de 2001 copiada en un documento Word, con unas 69.000 palabras). En el texto aparecen 32 errores ortográficos reales., que muestra la amplia cobertura de la base léxica de STILUS frente a la de Word XP. Los resultados se muestran en la tabla 2.

Tabla 2. Evaluación de la revisión ortográfica

REVISIÓN ORTOGRÁFICA	STILUS	Word XP
Total de alarmas generadas	175	768
Errores reales encontrados	32	32
Nombres propios no reconocidos	99	624
Palabras comunes no reconocidas	18	86
Palabras extranjeras no reconocidas	26	26

En cuanto al tiempo de procesamiento del sistema para las operaciones de segmentación y revisión ortográfica, es superior a 3500 palabras/seg, en un Pentium III a 800 MHz y 128 MB de memoria, con Windows 2000 Professional.

8.2 Revisión Gramatical

En cuanto a la revisión gramatical, los tiempos de procesamiento (en las mismas condiciones que antes) son de unas 1.500 palabras/seg.

La tabla siguiente muestra una tabla similar a la anterior. El número real de errores gramaticales en el texto es de 19 errores.

Tabla 3. Evaluación de la revisión gramatical

REVISIÓN GRAMATICAL	STILUS	Word XP
Total de alarmas generadas	20	26
Errores reales detectados	16	3
Alarmas falsas	4	23

9 Conclusiones y Trabajos Futuros

En este documento se ha presentado la descripción general y la arquitectura de STILUS, un sistema de revisión lingüística orientado a textos en castellano, mostrando el diagrama de bloques y describiendo cada uno de los componentes que permiten llevar a cabo los diferentes procesos sobre el texto, desde las operaciones básicas de filtrado y segmentación hasta la detección de errores gramaticales, ortográficos y de estilo. Se incluye también una tabla resumen de la evaluación del sistema frente al revisor más conocido y extendido en el mercado.

En cuanto a la revisión gramatical, es importante destacar el compromiso entre *precisión* y *cobertura*. Conseguir una precisión próxima al 100% tiene como inconveniente que la cobertura se reduce drásticamente, pues las reglas se centran en fenómenos muy particulares. En estos sistemas es igual de negativo no dar falsas alarmas dejando de detectar un gran número de errores, como aumentar la cobertura a costa de la fiabilidad.

De manera continua se trabaja en la actualización y expansión de los recursos de la base léxica, incluyendo nuevos términos generales, especializados en distintos campos o propios de las variedades del español de los países iberoamericanos. Por otra parte se trabaja en ampliación de la cobertura de las reglas gramaticales y de estilo, incrementando la base de reglas para detectar un número creciente de fenómenos lingüísticos. Además, debido a la comercialización de STILUS en EE. UU. (a través del portal http://www.spanishedit.com), se está trabajando en el desarrollo de nuevas reglas gramaticales específicas para la detección de errores habituales cometidos en el español de esa zona.

Por último, señalar que la detección de estructuras sintácticas en un idioma determinado es la herramienta básica de muchas otras facetas de la ingeniería lingüística, como la traducción automática o la generación de resúmenes, en los que también se están empleando los módulos aquí presentados.

Bibliografía y Referencias

[1] Agencia EFE. 1995. *Manual de Español Urgente.* Cátedra, Madrid.

[2] Alarcos Llorach, Emilio. 1994. *Grámatica de la Lengua Española*. Espasa Calpe, Madrid.

[3] El País. 1996. *Libro de Estilo El País.* Ediciones El País, Madrid.

[4] Goñi J.M., González J.C. and Moreno A. 1997. ARIES: A lexical platform for engineering spanish processing tools. *Natural Language Engineering Journal*, 3(4):317-345. Cambridge University Press.

[5] Iglesias C.A., González J.C., Goñi J.M., López J., y Nieto A. 1995. Procesamiento Semántico en la Arquitectura ARIES. *Novática*, (112-113):91-96.

[6] Marcos Marín, F. y España Ramírez, P. 2001. *Guía de gramática de la lengua española.* Espasa, Madrid.

[7] Martínez de Sousa, J. 1996. *Diccionario de usos y dudas del español actual.* Biblograf, Barcelona.

[8] Moreno Sandoval, M. 2001. *Gramáticas de unificación y rasgos.* Antonio Machado Libros, Madrid.

[9] Seco, M., Andrés, O. y Ramos, G. 2000. *Diccionario abreviado del español actual*. Aguilar, Barcelona.

[10] Seco, M. 1998. *Diccionario de dudas y dificultades de la lengua española*. Espasa Calpe, Madrid.

[11] Ramírez Bustamante, F., Sánchez León F. y Declerck T. (1998). «CONTEXT. Un corrector gramatical de bajo nivel». *Actas de la Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN*.

[12] Real Academia Española. 2001. *Diccionario de la Lengua Española.* Espasa Calpe, Madrid.

[13] Real Academia Española. 1999. *Ortografía de la Lengua Española.* Espasa Calpe, Madrid.