GRUPO 6: ANÁLISIS DE CORPUS

PRESIDENTE: JOSEBA ABAITUA

Morfeo: analizador morfológico y "tagger" del español

R. Pérez, D. Trotzig, X. Lloré

EUSLEM: un lematizador/etiquetador de textos en euskara

I. Aldezábal, I. Alegría, X. Artola, A. Díaz, N. Ezeiza, K. Gojenola

MORFEO:
ANALIZADOR MORFOLÓGICO Y 'TAGGER' DEL ESPAÑOL

Ricard Pérez, David Trotzig, Xavier Lloré
Departamento de I+D/NLU Micro Focus España SA

RESUMEN

Este artículo describe la herramienta de desambiguación categorial (MANÁ) que incorpora MORFEO, la plataforma de análisis morfológico desarrollada por ISS, actualmente Micro Focus España, en el marco del proyecto MMI^2 (ESPRIT P2474). Se exponen asimismo, las características principales del analizador (MORFO) y del diccionario general de la lengua española (SPAL) que soporta.

MORFEO/SPAL constituye una plataforma que, además del procesamiento de la morfología flexiva, incorpora el tratamiento de los fenómenos derivacionales más productivos y permite el reconocimiento de tiempos compuestos y lexías (con palabras flexionadas y términos incrustados). MANÁ es una herramienta modular de desambiguación categorial sobre la salida de MORFO, articulada sobre un filtro markoviano de orden 2 mejorado con un conjunto de reglas de reasignación de valores y con el recurso a un inventario de todas las tripletas posibles del español, complementos con los que MORFEO/SPAL consigue una precisión cercana al 100% en tratamiento de texto libre, con tablas construidas con volúmenes de muestra relativamente pequeños.

1. SPAL

El diccionario SPAL contiene 63174 palabras estándar, desarrolladas sobre 66711 raíces, que combinadas con 61 modelos flexivos y derivacionales expandidos con validaciones permite el reconocimiento de alrededor de 1160000 formas del español [1].

1.1. ESTRUCTURAS INFORMATIVAS

La Base de Datos de SPAL tiene un registro para cada raíz, con los siguientes campos: a) raíz primaria, b) acentuación, c) palabra estándar, d) raíz secundaria, e) categoría morfológica, f) categoría distribucional, g) modelo flexivo y/o derivativo, h) códigos morfosintácticos, i) divisibilidad/ indivisibilidad, k) dominio

[1] Para cada verbo regular se han considerado en el cómputo 116 formas, ya que los tiempos compuestos son reconocidos como unidades por el analizador. El tamaño del diccionario de Lengua Española SPAL, al cierre de l presente artículo contiene ya más de 80000 palabras estándar o lemas.

semántico y l) campos libres correspondientes a parámetros a determinar según el dominio de la aplicación.

Incluimos una breve referencia a algunos de los campos:

1.- (a) .- La descripción de la raíz se desarrolla mediante un lenguaje especializado (LENGCOM) que permite especificar lexías con más de una palabra flexionada, términos incrustados y marcar operaciones sobre éstos.

2.- (b) .- La existencia de un parámetro separado de la raíz para marcar la posición del acento gráfico permite el reconocimiento de formas independientemente de su correcta acentuación. El acento es tratado como una restricción respecto a los posibles análisis de una forma. Una palabra correctamente acentuada recibirá tan sólo el análisis correspondiente al término que especifica el acento en esa posición. Pero una forma sin acento o mal acentuada será analizada también como una ocurrencia de las entradas de diccionario que corresponden a su secuencia de letras, tanto si especifican acento como si no lo especifican

3.- (d) .- Las distintas raíces de una palabra forman una cadena de entradas en las que la raíz secundaria de una corresponde a la raíz primaria de la siguiente, hasta la última entrada, cuya raíz secundaria es 0.

4.- (f) .- Las categorías distribucionales, en número de 52, son la base de la estructura de datos utilizada por el filtro. Se han establecido partiendo de las categorías morfosintácticas habituales, que se han dividido en subgrupos según criterios de índole propiamente distribucional.

5.- (h).- En los códigos morfosintácticos se ha incluido información sobre las preposiciones regidas por los verbos auxiliares, el tipo de verbo pronominal, el valor reflexivo de los clíticos, etc, información que, independientemente de la especialización léxica propia de dominios particulares, es relevante para llevar a cabo ulteriores procesos de desambiguación contextual.

6.- (i) .- Los códigos de divisibilidad de una entrada de diccionario indican si, dado el análisis de una secuencia respecto de esa entrada, el analizador debe considerar tan sólo ese análisis (indivisibilidad) o si deben buscarse también otros posibles análisis alternativos (divisibilidad). En el caso de lexías permiten gestionar si se aceptan múltiples análisis: el de la lexía como unidad y el correspondiente a la secuencia de sus componentes, o si tan sólo cabe el primero de ellos. Cuando se trata de palabras simples controlan si la combinación raíz-sufijo indicada en esa entrada es decisiva o se permiten otros análisis.

por ejemplo: "generales" -> 'imperativo de generar' + enclítico 'les' , junto con el plural de "general" como nombre y como adjetivo.

1.2. ESTRUCTURA DE LOS DATOS

El diccionario de raíces (DICO) está organizado con una estructura TRIE. Esta es una estructura de árbol, donde cada rama: secuencia de caracteres alfabéticos, incluyendo el blanco, desarrolla una raíz - clave primaria para la indexación de los registros. Este formato permite que la identificación de la raíz en la B.D. sea paralela al proceso de fragmentación de izquierda a derecha del texto que se analiza. De este modo el reconocimiento morfológico no está guiado por el espacio en blanco, con lo que se optimiza la identificación de lexías. Asimismo, supone una importante economía de espacio respecto a la estructura habitual de otras bases de datos ya que la secuencia de caracteres comunes de todas las raíces, empezando por la izquierda, es recogida una única vez.

2. MORFO

En este apartado se expone el algoritmo del analizador morfológico y el lenguaje de especificación de raíces. La posibilidad de especificar 'palabras' formadas por múltiples términos no tiene tan sólo relevancia desde el punto de vista lexicográfico, sino que también es importante para un correcto funcionamiento del filtro de desambiguación. La asignación de la categoría correcta a una lexía indivisible produce resultados más ajustados que si se accediera a las tablas de frecuencias con la secuencia de las categorías correspondientes a cada uno de sus componentes.

2.1. LENGCOM

El lenguaje de especificación de raíces LENGCOM permite describir las palabras simples de manera muy sencilla y también lexías. Las lexías pueden estar compuestas por términos con o sin flexión y también aparecer con términos incrustados sobre los que son posibles operaciones de anteposición o posposición; LENGCOM permite indicar la flexión principal y las secundarias, así como caracterizar como opcionales u obligatorios los términos que las componen. En detalle, algunas de sus funcionalidades son:

- Marcas sobre los sufijos. Diferencian entre la raíz cuyo sufijo, 'activo' , aportará sus rasgos al análisis final de la lexía, y aquellas que flexionan, pero sin aportar rasgos al análisis de toda la unidad.

- Descripción sintáctica de la lexía. La descripción de la lexía se desarrolla con cuatro tipos de información:

a) Elementos que la constituyen. Pueden especificarse por su categoría morfológica, forma superficial, palabra estándar y rasgos morfosintácticos.

b) Códigos de opcionalidad, obligatoriedad o prohibición sobre los elementos componentes e incrustados.

c) Operaciones booleanas sobre los rasgos descriptivos de estos términos.

d) Operaciones sobre los elementos: integración, desaparición, anteposición, posposición.

2.2 ANÁLISIS MORFOLÓGICO.

Exponemos una descripción no formalizada del algoritmo del analizador morfológico para el caso más simple, cuando no se trata con lexías.

La función principal es la búsqueda de secuencias en el árbol de raíces cada vez incrementadas en un carácter, correspondientes a la fragmentación por la izquierda de la entrada. Cuando se encuentra el final de una rama en el árbol, se va a buscar la información sobre los modelos de flexión de esa raíz.

Se procede finalmente a comprobar que el resto de la entrada hasta el espacio blanco corresponda a uno de los sufijos de esos modelos, partiendo siempre de las fragmentaciones de la palabra que toman la raíz de mayor longitud de caracteres. La estructura de información resultante del análisis morfológico combinará la que recoge el diccionario de raíces con la aportada por el sufijo. Si los códigos de divisibilidad lo permiten, se buscarán todos los posibles análisis raíz-sufijo, si no, el proceso terminará con el primero de ellos.

Procedimientos especializados tratan las palabras con mayúscula, los tiempos compuestos de los verbos y los enclíticos, las posibles combinaciones entre los cuales han de ser consideradas durante el análisis de verbos. Un aspecto relevante es el desarrollo de modelaciones adecuadas para aquellos casos en los que la concatenación con el enclítico provoca modificaciones en la raíz ("dad" / "daos", "movamos"/ "movámonos"). Para optimizar el tratamiento de enclíticos se ha implementado un 'switch' que habilita o deshabilita al analizador para reconocer enclíticos en los tiempos personales. En su posición habitual no permite reconocer formas como " encontréme" o "hallábase", tan sólo soporta enclíticos en imperativo, presente de subjuntivo y formas no personales. Si se desactiva, se obtiene los análisis verbo-enclítico correspondientes a cualquiera de los tiempos verbales. Así, en esta modalidad de análisis encontraremos

que la forma "amante" obtiene análisis de presente de indicativo 'aman' + clítico 'te', además de los usuales de nombre y adjetivo.

3. MANÁ

Como ya se ha expuesto, bajo este nombre se identifica el módulo de desambiguación categorial, articulado alrededor de un filtro markoviano, posterior al analizador morfológico MORFO.

3.1. PROCESOS DE DESAMBIGUACIÓN PREVIOS AL FILTRO

Algunos procesos morfológicos y morfográficos llevan a cabo ya una acción de desambiguación al ofrecer como resultado una única alternativa para formas que podrían dar lugar a ambigüedades. Cabe citar entre ellos el tratamiento de las contracciones ("del" -> 'de - preposición'+ 'el - artículo'), de los enclíticos ("cércalas" -> 'imperativo de cercar'+ 'clítico acusat. fem. plu.', no se acepta el análisis como adverbio ni nombre más artículo), las lexías indivisibles ("a buen paso" - > 'adverbio: rápidamente'), los tiempos compuestos ("había extraviado" -> 'pluscuamperfecto de extraviar') y, finalmente, la acentuación ("intérprete" obtendrá analisis como nombre a despecho de la existencia en el diccionario para las formas "interprete" o "interpreté").

3.2. DESCRIPCIÓN DEL FILTRO

La entrada de MANÁ es la salida del analizador morfológico (MORFO), que ofrece entre cada par de posiciones de la secuencia de entrada todos los posibles análisis resultantes de la aplicación de los procesos y criterios reseñados.

La función de MANÁ es asignar a cada subsecuencia de la lista de entrada para la que haya análisis con categorías distribucionales alternativas una única cadena de estructuras morfológicas. MANÁ no considera información léxica y, en principio, no le llegan ambigüedades correspondientes a homonimias de tipo léxico si no tienen reflejo en características propiamente morfológicas. Efectivamente, las entradas del diccionario morfológico no discriminan entre acepciones para una misma palabra, excepto en los casos en los que la oposición de significado es correlativa con una oposición de modelación ('modelo' - masc. / 'modelo' - fem. y masc.). Salvo estos casos y los dobletes de raíces correspondientes a verbos pronominales y no pronominales, o a adjetivos y nombres deverbales e idiosincráticos, las ambigüedades

que pasan a MANÁ siempre implican oposición de categoría: 'dado' (nombre) / 'dado' (participio), 'sal' (nombre) / 'sal' (verbo).

La base del filtro es un modelo markoviano de orden 2 que recoge la probabilidad de transición para cada secuencia de tres categorías. La operación principal consiste en consultar la tabla de frecuencias, elaborada a partir del análisis de muestras, y calcular la probabilidad de cada uno de los caminos alternativos entre puntos sin ambigüedad. Este cálculo se lleva a cabo mediante la suma de las probabilidades de sus tripletas cuando en todos los caminos hay el mismo numero de categorías, y mediante la media de esas probabilidades cuando el número de tripletas en los caminos es distinto. Esto ocurre en particular cuando compite el análisis de una lexía como unidad contra el de la secuencia de sus palabras.

Este procedimiento básico comporta operaciones más complejas en algunos casos que se expondrán más adelante.

En los casos citados, en que compiten alternativas con la misma categoría distribucional, el filtro no puede resolver la ambigüedad, que deberá ser tratada por operaciones de desambiguación contextual a la salida de MORFEO. Estas operaciones están implementadas en un módulo de 'microsintaxis' compuesto por un pequeño conjunto de reglas operativas sobre los contextos más o menos cercanos al punto de la ambigüedad. El presente artículo no se extiende en la descripción de estas reglas ni en su forma de aplicación.

3.2.1 TABLAS DE FRECUENCIAS

La elaboración de las tablas con las frecuencias de las tripletas de una muestra es un proceso que se realiza sobre la salida que ofrece MORFO para ese corpus. En el entorno de trabajo desarrollado alrededor de MORFEO se han implementado dos herramientas para automatizar este proceso, mediante dos estrategias alternativas:

a) Desambiguación durante el análisis morfólogico. Durante el proceso de análisis morfológico un programa interactúa con el lexicógrafo, solicitando su decisión siempre que se presenten análisis alternativos. Con el resultado de este análisis supervisado se construyen automáticamente las tablas.

b) Desambiguación semiautomática. Para construir tablas específicas de un dominio, disponiendo ya de tablas de uso general para narrativa, hemos desarrollado un proceso de corrección manual sobre la salida de MORFEO, trabajando con las tablas generales. La intervención del lexicógrafo es mínima, corrigiendo

tan sólo aquellos casos en los que el filtro de uso general haya fallado. Un programa construye automáticamente las tablas a partir del fichero con el análisis correctamente desambiguado.

3.2.2 PROCESO DE FILTRAJE

Como ya se ha expuesto, el proceso básico de desambiguación consiste en el cálculo de la probabilidad de cada uno de los caminos alternativos para cada subcadena de estructuras morfológicas limitada por delante y por detrás por 2 elementos no ambiguos, unidad de tratamiento que denominamos 'vientre'.

A fin de superar la explosión combinatoria asociada al alto número de ambigüedades que presentan algunas oraciones, el principio de expansión de los vientres se hace de forma inteligente, desechando dinámicamente caminos alternativos de menor probabilidad de acuerdo con la expansión necesaria para acoger el número de alternativas de la palabra siguiente en correspondencia con el espacio máximo de registros del campo de soluciones, parámetro que el usuario puede determinar externamente. Cuanto más alto sea éste, más lento será el proceso de filtraje. Un límite máximo evita los fallos por limitación física de la memoria.

3.2.3. PROBABILIDAD DE COOCURRENCIA DE CATEGORÍAS

En un filtro markoviano estándar, la probabilidad de una tripleta - $Pr(c_1,c_2,c_3)$ - depende directamente del número de apariciones de esa secuencia de categorías $[c_1,c_2,c_3]$ en la muestra usada para construir la tabla de frecuencias.

Cuando se utiliza un filtro markoviano para desambiguar texto y los símbolos 'c_i' de la tabla de transiciones corresponden a conjuntos de palabras, la categoría asignada a una palabra ambigua entre dos categorías alternativas es directamente dependiente de la extensión de los elementos del idioma incluidos en cada una de las categorías alternativas. De ahí que, cuando aparecen varias categorías en un mismo contexto, las tablas ofrezcan un valor más alto para aquella tripleta que incorpora la categoría más poblada o extensa de entre las alternativas a desambiguar.

En nuestra opinión, cuando se trabaja con categorías lingüísticas como estados de un modelo markoviano, en el momento de decidir sobre la asignación de un término a una u otra categoría no debe considerarse únicamente el número de veces que cada una de esas categorías ha sido registrada en el contexto de que se trate, sino también la probabilidad del término de pertenecer a una u otra de las categorías.

En nuestro filtro, con el objeto de poder reflejar la estructura del sintagma nominal, se han establecido categorías distribucionales para cada uno de los distintos tipos de determinantes. El término "unas" es ambiguo entre verbo (v) y determinante indefinido (V). Si tratásemos en el proceso de filtraje únicamente con las frecuencias que corresponden a las apariciones de las categorías en la muestra, la gran mayoría de ocurrencias de ["unas" + nombre] serían resueltas asignando a "unas" categoría 'v', ya que en las muestras hay muchas más apariciones de verbo ante nombre, que no de determinante indefinido ante nombre. Ahora bien, si combinamos este dato con la probabilidad de "unas" de ser verbo y de ser determinante indefinido, que es inversamente proporcional al número de palabras en cada una de esas categorías, el resultado del filtraje compensará ese efecto de desproporción.

A fin de disponer de un valor numérico para cada tripleta que refleje esta doble perspectiva se calcula el valor de cada vientre como la suma de la frecuencia relativa de cada una de sus tripletas, desarrollada como la probabilidad de coocurrencia de sus categorías:

$$\log_2 \frac{Pr\,(c_1, c_2, c_3)}{Pr(c_1)\,Pr(c_2)\,Pr(c_3)}$$

3.3. ESTRATEGIAS DE CORRECCIÓN

El uso de tablas de frecuencias, obtenidas del análisis de un corpus necesariamente reducido, en un filtro markoviano para tratar texto libre comporta algunos desajustes para cuya resolución se han diseñado estrategias particulares:

a) Presencia en el texto tratatado de tripletas no registradas. Si se trabaja con el cálculo de frecuencias estándar, la probabilidad de transición asociada a una de estas tripletas es 0. Si se usa la fórmula de probabilidad relativa expuesta, su valor será $\log_2(0/0)$ o $\log_2(0)$. Tales valores impiden un procesamiento satisfactorio.

b) El cálculo adoptado prima la asignación de una palabra ambigua a la categoría menos extensa atestada en el contexto en cuestión. Si en la muestra hay una única aparición de una categoría (c_1) , siempre que en su contexto aparezca un término que es ambiguo entre esa categoría y otras que tenga atestiguadas apariciones en otros contextos, le será asignada 'c_1'. Esto parece implicar que las tablas deben proceder del análisis de muestras muy grandes.

c) No parece existir ningún criterio seguro para establecer las categorías 'distribucionales' sobre las cuales trabaja el filtro. Por lo general se sigue un proceso de prueba y error sobre la base de las categorías morfosintácticas tradicionales. Obviamente, los resultados que ofrece el filtraje dependen de la categorización adoptada, cuantas más categorías se adopten, más precisos serán los resultados, pero también, más complejo el manejo de su explosión combinatoria.

d) La presencia de palabras desconocidas constituye un problema para el proceso de filtraje, no tienen categoría y no tendría sentido incluir en las tablas una categoría que correspondiese a la distribución hipotética de un concepto tal como "palabra desconocida".

3.3.1. ESTRATEGIA DE LAS TRIPLETAS POSIBLES

Para tratar la presencia en el texto de tripletas no registradas en la muestra se ha construido una base de datos con todas las tripletas posibles del español. Esto permite desambiguar texto libre con tablas construidas sobre una muestra pequeña, sin tener que abortar el procesamiento por la ausencia de una tripleta dada.

Cuando el filtro trabaja con estas tripletas es posible desambiguar palabras en contextos no atestiguados en la muestra., rechazando algunos caminos como no gramaticales y considerando el valor de otros que sí corresponden a transiciones registradas como posibles, aunque no aparecen en la muestra.

Sobre la base de las 52 categorías adoptadas se obtienen 140608 combinaciones de tripletas, de las cuales se han seleccionado aproximadamente 55500 como posibles en la lengua española.

Para calcular la probabilidad de las tripletas de este grupo que no aparecen en la muestra se hace la media de los valores mínimos de las tripletas correspondientes a los tres contextos de dos categorías que recubren la tripleta de partida. Si la tripleta en cuestión es <A,B,C>, los valores a considerar en el cómputo son los mínimos que ofrecen los conjuntos de tripletas determinados por <A,B,*>, <A,*,C> y <*,B,C>. Cuando alguno de éstos es el conjunto vacío se asigna a la tripleta el valor mínimo de todas las tripletas de la muestra.

3.3.2. ESTRATEGIA DE REASIGNACIÓN DE VALORES

Con el fin de corregir los problemas derivados de la irregular distribución de las palabras en las categorías adoptadas, así como para compensar algunos de los efectos asociados al uso de pequeñas muestras y a la

inclusión de tripletas atestiguadas en general para la lengua, se ha confeccionado un conjunto de reglas de reasignación de valores a tripletas.

Un aspecto que tratan en particular estas reglas son los casos en que términos muy frecuentes son ambiguos entre dos categorías (c_1/c_2) y el filtro les asigna en ciertos contextos la categoría errónea. Si uno de estos términos es categorizado erróneamente, según las tablas procedentes de una muestra, en un contexto como 'c_2', pueden establecerse reglas que igualen o incrementen los valores de las tripletas a la izquierda o derecha de 'c_2' en ese contexto respecto los valores de las que incluyen a 'c_1' en lugar de 'c_2'. Serán los valores de las tripletas correspondientes al lado no modificado los que decidan la asignación.

El objetivo de las reglas de reasignación es expresar proporciones que deben mantenerse entre valores de tripletas para un lenguaje natural gramatical independientemente del texto de muestra a partir del cual se hayan elaborado las tablas.

Al usar tablas que resulten de la aplicación de estas reasignaciones a las tablas obtenidas de una muestra, MANÁ queda configurado como herramienta independiente de esa muestra, manteniendo un conjunto constante de relaciones entre las tripletas de la lengua como criterio general de desambiguación. En el apartado sobre evaluación se incluye un análisis de su relevancia respecto la eficacia global del filtro.

El núcleo de proporciones entre valores distribucionales de las distintas categorías que resulta de aplicar las reglas permite generalizar el uso de MANÁ para distintos estilos y dominios lingüísticos gramaticales o bien formados del español.

3.3.3. ESTRATEGIA PARA LAS PALABRAS DESCONOCIDAS

Para abordar la asignación de categoría a las palabras desconocidas se ha desarrollado un proceso inteligente que considera los procesos flexivos y derivacionales del español, así como los de otras lenguas originarias de barbarismos. Téngase en cuenta que este proceso enfoca sobre todo el fenómeno de la innovación lingüística más que los fenómenos de error y equivocación en la escritura de los textos. La prevalencia de este enfoque, sin embargo, no entra en contradicción ni se opone a otros procesos que pudieran crearse para enfocar el otro objetivo.

El proceso implementa tres criterios:

a) Una palabra desconocida tan sólo podrá pertenecer a una clase abierta: nombre, adjetivo, adverbio o verbo.

b) Ante tripletas alternativas para el contexto en que se encuentra una palabra desconocida, si están en conflicto tripletas recogidas de la muestra con otras tan sólo atestiguadas como posibles, se pondera más la asignación correspondiente a las tripletas de la muestra.

c) La probabilidad de que una palabra pertenezca a una categoría se hace depender de su sufijo. A los distintos afijos se les ha asignado un factor por el que se multiplica esa probabilidad. El sufijo 'mente' multiplica por 5 la probabilidad de que la palabra desconocida sea adverbio, el sufijo 'ción' multiplica por 4 la probabilidad de que sea nombre.

El procedimiento para asignar categoría a una palabra desconocida 'x' en el contexto <A,x,B> recupera los valores de todas las tripletas resultantes de instanciar 'x' a las categorías de las clases abiertas citadas y lleva a cabo los cálculos correspondientes a los criterios b) y c).

Estos procedimientos aseguran, como se muestra en la evaluación, un alto grado de precisión en la asignación de categoría a aquellas palabras que corresponden a los paradigmas flexivos de la lengua, tanto neologismos como términos especializados.

3.3.4. CONFIGURACIÓN

La actuación de MANÁ queda caracterizada por las tablas de frecuencias a partir de las cuales tiene lugar el cómputo de valor para cada uno de los caminos alternativos de un 'vientre'.

La composición de las tablas se configura para cada aplicación con un programa de carga especificando tres parámetros: a) fichero base de tripletas extraídas de una muestra, b) integración o no del fichero de tripletas posibles y c) aplicación o no de las reglas de reasignación de valores.

Tal como se mostrará en el siguiente apartado, las tablas resultantes de aplicar positivamente las opciones b) y c) a las tablas extraídas de una muestra de 30000 bytes de un libro técnico sobre redes de comunicación aportan a MANÁ un alto grado de eficacia para distintos estilos de lenguaje. Consideramos que MANÁ, con estas tablas, puede ya ser integrado óptimamente en múltiples aplicaciones. No obstante, está a nuestra disposición la posibilidad de otras configuraciones que pudieran, por ejemplo, tener como base tablas procedentes de muestras de estilos de lenguaje muy específicos o característicos.

4. EVALUACIÓN

La evaluación de resultados refleja el rendimiento de MORFEO/SPAL, donde se integra el módulo de filtraje MANÁ, y paralelamente resulta ilustrativa sobre el volumen de muestra óptimo para un filtro con las características del descrito.

4.1. CORRELACIÓN ENTRE VOLUMEN DE MUESTRA Y PRECISIÓN

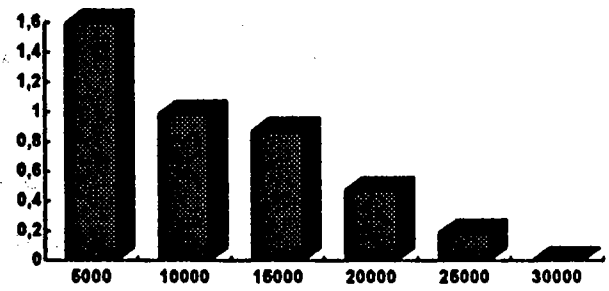
Para llevar a cabo este experimento se han construido tablas base procedentes de volúmenes crecientes de texto : 5000, 10000, 15000, 20000, 25000 y 30000 bytes de un libro técnico sobre redes y se ha comprobado la exactitud del filtro alimentado con cada una de estas tablas al analizar un volumen de 30000 bytes distintos del mismo libro. En todos los casos las tablas de trabajo han resultado de la combinación de las tablas base con el fichero de tripletas posibles y la aplicación de las reglas de reasignación de valores.

- PORCENTAJE DE ERRORES RESPECTO AL NÚMERO TOTAL DE PALABRAS -

Estos datos recogen el porcentaje de asignaciones erróneas respecto al número total de palabras del texto, en relación con los volúmenes de muestra utilizados al confeccionar las tablas base.

Al considerar la tarea que lleva a cabo el filtro es relevante tener en cuenta el elevado porcentaje de palabras ambiguas que aparecen en la salida de MORFEO. Este porcentaje se sitúa, en el volumen de 30000 bytes del libros de redes sobre el que se ha realizado la experiencia, en un 40,1405 %. Este elevado porcentaje es, en parte, la contrapartida a la robustez del sistema al no tratar como discriminante la ausencia de acento. Todas las ocurrencias de "de" resultan ambiguas entre el verbo 'dar' y la preposición, "el" es ambiguo entre artículo y pronombre, "que" lo es entre relativo, interrogativo y conjunción. También incide en este porcentaje la fina discriminación o granularidad morfosintáctica recogida en el diccionario: distintas entradas para nombres y adjetivos con las mismas formas ("negro", "físico", "mecánico"), distinciones entre las formas pertenecientes a las categorías cerradas ("que" - relativo/ conjunción, "aquel" - determinante/ pronombre), etc.

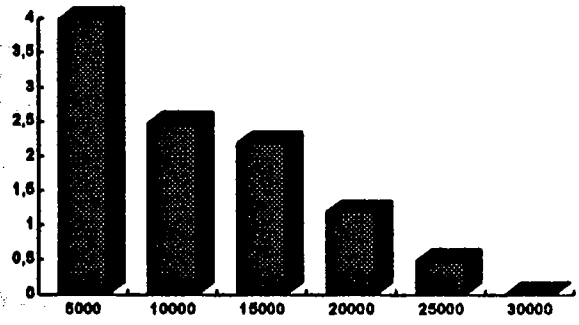
VOLUMEN DE LA MUESTRA (BYTES DE TEXTO)	5000	10000	15000	20000	25000	30000
%ERRORES	1,5957	0,9878	0,8739	0,4749	0,1900	0,0000



- PORCENTAJE DE ERRORES RESPECTO AL NÚMERO DE PALABRAS AMBIGUAS -

La siguiente tabla ofrece los porcentajes calculados sobre el número total de palabras desambiguadas

VOLUMEN DE LA MUESTRA (BYTES DE TEXTO)	5000	10000	15000	20000	25000	30000
%ERRORES	3,9754	2,4610	2,1770	1,1832	0,4733	0,0000



Al poner en correlación el número de errores con los volúmenes de muestra se observa un punto de inflexión entre los 15000 y 20000 bytes, a partir del cual la curva tiende a ser asintótica respecto al eje de abscisas (BYTES DE TEXTO), tanto si se aplica el filtro aumentado con las reglas de reasignación, como si se aplica sin ellas. Esto hace prever que los beneficios con tablas obtenidas de muestras mayores de texto serán menos que proporcionales en relación al esfuerzo empleado, llegando a ser ínfimos a partir de los 30000 caracteres de muestra.

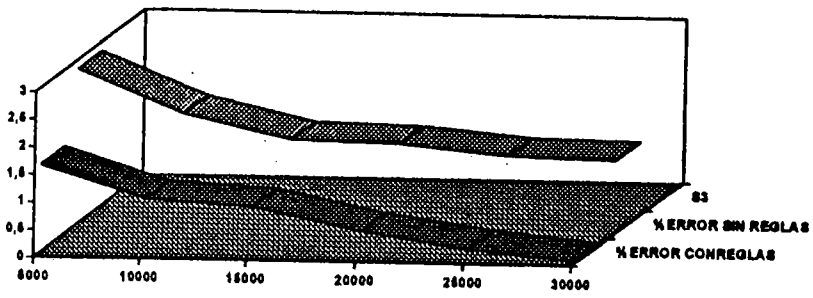
4.2. REGLAS DE REASIGNACIÓN

A fin de evaluar el peso de las reglas de reasignación de valores en la eficacia del filtro se han llevado a cabo mediciones de los resultados obtenidos por MANÁ dotado de datos puramente estadísticos: tablas de las muestras (5000, 15000 y 30000 bytes), más el fichero de tripletas posibles. La comparación de estos resultados con los que se obtienen al aplicar las reglas revela la importancia de esta estrategia. El 'tagger' implementado en MANÁ se configura como un modelo mixto: base estadística + reglas de reasignación.

El conjunto resultante , sobre la base de muestras reducidas, alcanza un grado de eficacia y una capacidad de generalización a textos de estilos similares que difícilmente se hubieran podido obtener mediante recursos puramente estadísticos.

PORCENTAJE DE ERRORES RESPECTO NÚMERO TOTAL DE PALABRAS

VOLUMEN DE LA MUESTRA (BYTES DE TEXTO)	5000	10000	15000	20000	25000	30000
% ERRORES SIN REGLAS	2,8305		1,5957			1,3297
% ERRORES CON REGLAS	1,5957	0,9878	0,8739	0,4749	0,1900	0,0000



4.3. EXPORTACIÓN DE TABLAS

Al usar MORFEO, con MANÁ alimentado por las tablas resultantes de combinar el fichero de tripletas posibles con las obtenidas de la muestra de 30000 bytes del libro de redes y aplicar las reglas de reasignación, para analizar texto periodístico, 5909 palabras, se han obtenido tan sólo 18 errores. Esto corresponde a un porcentaje de aciertos del 99,6953 % respecto el número total de palabras.

Estos resultados permiten corroborar que MORFEO, con MANÁ configurado con las opciones citadas, puede alcanzar resultados óptimos con tablas provenientes de muestras pequeñas al ser aplicado a textos de distintos estilos.

4.4. ASIGNACIÓN DE CATEGORÍA A LAS PALABRAS DESCONOCIDAS

La evaluación de las asignaciones de categoría a las palabras desconocidas del texto periodístico nos ofrece el siguiente resultado: de 36 asignaciones MORFEO ha cometido tan sólo 1 error. Esto corresponde a un acierto del 99,9722 %.

4.5. VELOCIDAD DE PROCESO

El procesamiento de 30000 bytes se ha realizado en 9 minutos y 44 segundos sobre una Sparcstation 10 con 8 MB de RAM. Esto representa aproximadamente un promedio de 1,63 páginas x minuto

5. CONCLUSIÓN

La plataforma MORFEO formada por SPAL, MORFO y MANÁ constituye un módulo adaptable a idiomas, estilos y aplicaciones diversos. Esta configuración con las tablas de lengua española expuestas es aplicable industrialmente a cualquier texto que precise análisis morfológico con desambiguación. Puede partirse de la base léxica SPAL para desarrollar diccionarios especializados apoyándose en la estructura semántica correspondiente a la información codificada en el diccionario.

La metodología MORFEO puede aplicarse a otras lenguas modificando tan sólo los datos del diccionario de raíces, los modelos de sufijos y el inventario de tripletas. Asimismo, los recursos de construcción de tablas permiten desarrollar fácilmente configuraciones para estilos de texto particulares o ceder al usuario estas herramientas constructivas. Igualmente, la capacidad de intervenir en la lematización incorporada en SPAL, posibilita implementar en el diccionario mismo estrategias de normalización por convergencia conceptual adecuadas en dominios semánticos específicos, con lo que el proceso de MORFEO ofrece una salida de uso directo en procesos de recuperación de información con criterios de articulación conceptual.

Como se ha citado al inicio del texto, las herramientas aquí descritas han sido desarrolladas en el marco del proyecto MMI^2 (ESPRIT P2474), cuyo objetivo ha sido el desarrollo de un prototipo de interfaz multimodal para la comunicación hombre-máquina. Integrado en el modo español de interacción en lenguaje natural del interfaz, MORFEO, con tablas de filtraje y un diccionario especializados para la aplicación, ha constatado su rendimiento y efectividad . El modo español procesa 200 palabras en menos de 3 segundos y medio para llegar a la representación lógica del significado.

Por tanto, MORFEO dispone en la actualidad de un ámbito aplicativo general sobre las distintas tecnologías del texto: indexación, sistemas de traducción automática, comunicación hombre-máquina, recuperación de información, corrección, acceso a bases de datos y de conocimiento, etc. MORFEO puede ser integrado mediante una sencilla instrucción embebida en el código de lenguajes de programación de alto nivel como COBOL, C, PROLOG, etc. Para cualquier entorno MORFEO ofrece como salida un texto desambiguado que configura un estándar unificador de posteriores tratamientos sean de tipo estadístico, sintáctico, semántico, terminológico o de gestión con recuperación de información.