Analisis morfologico automatico del espanol

Montserrat Meya
Munich

1 Objetivos
2 base de datos
3 componentes

1 Objetivos

El análisis morfológico de las palabras de un texto puede realizarse:

a) a nivel de cadena, las palabras son despojadas de 1,2,3,...n grafemas mediante algoritmos que asignan categorías, sin diccionario.
b) a nivel de palabra, bien con diccionario de formas de palabras, o bien con diccionario de lemas- formas base - . En este caso se trata fundamentalmente de búsqueda de diccionario reduciendose el número de algoritmos.
c) a nivel de morfema, con un diccionario mucho más reducido - hay muchos menos morfemas que palabras- pero con mayor aparato procedural.

Un análisis morfológico a nivel de morfema -tipo c- tiene la ventaja de:

- necesitar menos memoria para el almacenamiento de los datos lingüísticos -menor número de entradas-.
- poder relacionar entre sí diversas formas de palabras con una raíz común, aun en aquellos casos en que se trata de formas fuertes o irregulares.

El análisis morfológico aquí expuesto puede ofrecer:

- la descomposición de cada palabra de un texto escrito en sus morfemas respectivos
- la categorización morfológica
- lematización o reducción de las formas de palabras a su forma base(p.ej.: "trabajaba" ---->"trabajar")
- especificación de la raíz (p.ej.:"trabajaba"--->"trabaj")

El objetivo del analizador morfológico es obtener descriptores que den cuenta del tipo de información tanto de los documentos-textos almacenados- como de la demanda formulada por el usuario a un banco de datos.
El "recall" es obviamente enorme, si no se recurren a otras funciones de retrieval que precisen o ajusten la información a considerar.

2 Base de datos

El modelo de análisis morfológico automático es aplicable a cualquier lengua -está implemantado para el alemán,español e inglés- sólo la base de datos y pocos programas específicos varian según la lengua de que se trate.
La base de datos consta:

A : Listado de morfemas categorizados a partir de propiedades
- léxico-sintácticas (raíz léxica o no,nominal,verbal,...)
- flexivas : tipo de flexión,tipo de variante,...
- distributivas: reglas de serialización,posición en cadena..)

B : tabla de condiciones para las transformaciones(cuento---> contar; transformaciones ue --> o,ie-->e,g-->j,etc...)

C : Tabla de formas fuertes (alomorfos) a los que se hace referencia por un puntero.

D : Tabla de formantes flexivos con informacion detallada a usar en el posterior análisis sintáctico.

E : Palabras funcionales (gramaticales)- stop-words.

F : Tabla de reglas para la descomposición.

2.1 Especificacion subcategorizacion de los datos

En la cadena gráfica de un asiento pueden coincidir varios morfemas. Cada asiento es ,pues, un morfo.Por ej.
"libr" representa 4 morfemas : libro.libra.libre.librar
Cada entrada/asiento tiene asignado el siguiente patrón de 32 bits.

LEX	NOM	VRB	ADJ	ADV	NUM	FUN	GEN	Conj.		C	Trafo		ACT	D-O
Ø	o	a	e	o/a	Ø/a	ALO	DIF	AFF	PRE	ADn	FLX	SUF	Subcat	ADv

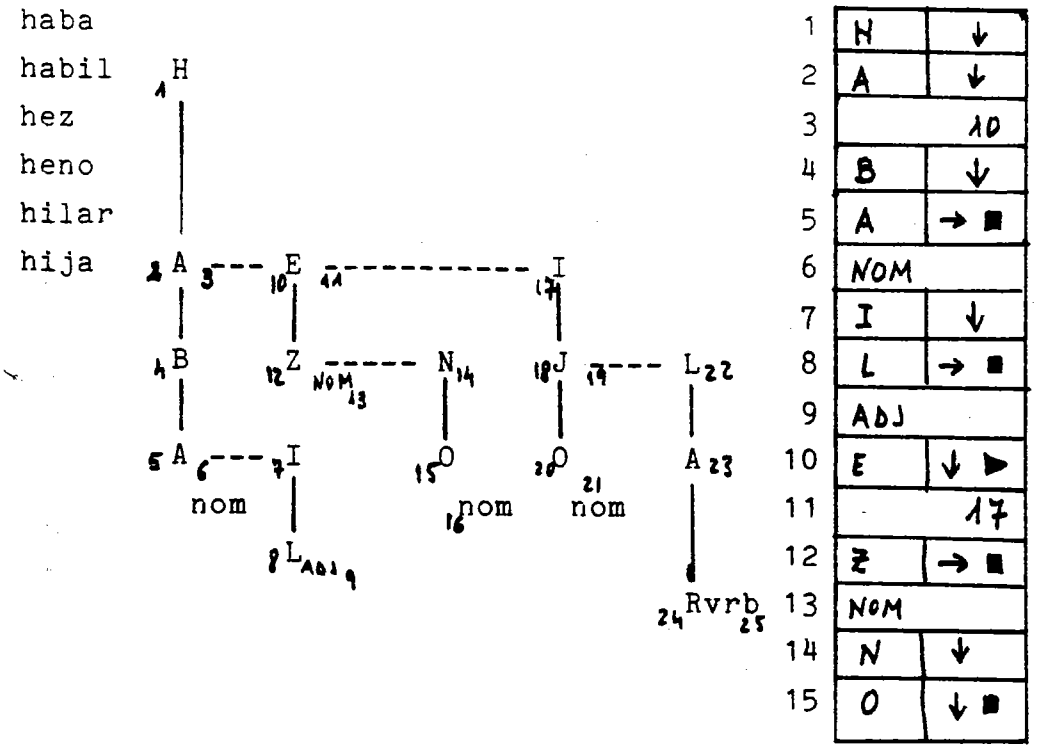
Categoriz.morfológica: tipo de morfema: nominal,verbal,etc, Subcategorizacion: GENero,Conjugacion(bits 9-10),tipo de flexión (bits 17-22),alomorfo,transformaciones posibles,SUFijo,Adnnominal,etc...
Los alomorfemas que por procesamiento pueden ser reducidos a su morfema originario no tienen asiento.Por ejemplo tiene asiento "conoc"-er, pero no "conozc-o" ni tampoco "cuent"-o que son transformados en el proceso de reconocimiento.
Ejemplo de asientos del diccionario:

```
ventan : LEX          NOM       a
cont   : LEX          NOM VRB o a  ar ue D-O
alguien: FUN
o      : FUN AFF FLX NOM VRB ADJ ADV NUM o o/a ar er ir
ción   : AFF SUF      NOM              0          Deverbal,....
```

La tabla de transformaciones es procesada por SPATZ. Esta tabla tiene almacenada en forma arborea las subcadenas que permiten transformacion junto con una categoria (p.ej:1=c-->z)

2.2. Almacenamiento de los datos

Las diversas listas estan comprimidas/almacenadas en forma arborescente lo que ademas facilita la búsqueda. Cada letra del alfabeto tiene un árbol que será más o menos extenso según el número de ramificaciones que tenga. El árbol para"p" es mayor que el de "h" porque hay muchos más morfemas que empiecen por tal letra. Se trata de una optimación de una representación árborea bisecuencial. Las siguientes palabras estarían almacenadas:

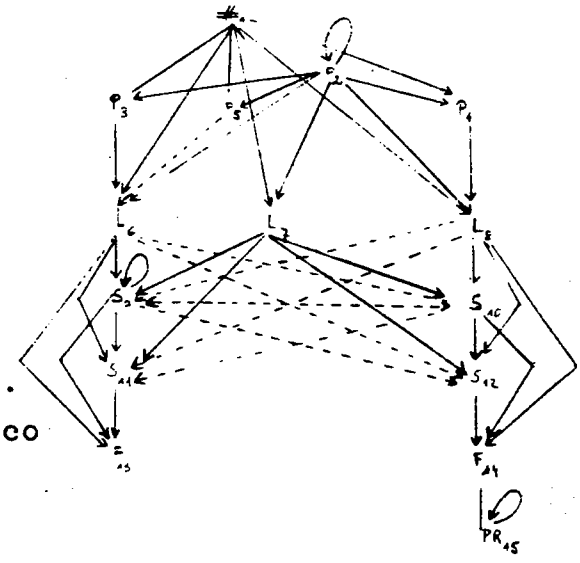


2.3 Tabla de reglas para la descomposicion

La descomposición de una palabra en sus morfemas se hace a partir de 15 posibles estados y las posibles transiciones entre estos estados:

1 =inicio 2= prefijo recursivo 3=prefijo adnominal
4= pref.adv. 5= prefejo denominal 6= morfema no verbal
7= morf.ligado 8= morfema verbal 9= suf.adnominal recursivo
10=suf.adverbal recursivo 11= suf. " no-recursivo
12=suf.adverbal no recursivo 13= flexion no verbal 14= pronombre

Esta red de transición esta expresada en reglas de producción del tipo:



premisa=complejo de posibles propiedades del morfema a tratar
condicion=propiedades exigidas simultaneamente para activar la acción.
acción=transicion a otro estado morfológico

Este saber operativo esta almacenado en bits -formato hexadecimal correspondientes a las propiedades morfológicas vistas arriba. Cada regla esta codificada como una estructura de datos separada e independiente. La regla siguiente permite la transicion de morfema léxico no verbal L6 a su flexión correspondiente

i	0101 1110 0000 0000 0010 0001 1011 0000
ii	0100 0000 0000 0000 0000 0001 1011 0000
iii	0013 0002

De un estado x se pasa a F13 (flexión no verbal) si los bits 2,24,25,27,28 están prendidos. En caso afirmativo se comprueba que F13 y el morfema anterior sean del mismo subtipo (la condicion 0002) ."ventan" exige "-a", y "as" es del tipo "a".Entonces se pasa al estado 13 o sea "as" recibe tal estado.

3 Componentes

Hay varios programas que:

1. preparaN y editan los datos : generan la codificación interna
   : comprimen los datos (árboles)

   A partir de estos formatos comunes para las 3 lenguas se pueden obtener listas comparativas de aspectos o categorías a partir de determinado vector. Estos serían productos secundarios a obtener de interes lingüístico.
2. Programas de entrada/salida según el proceso a adoptar a continuación. El formato del output será diferente según se quieran invertir o no los resultados lingüísticos, o se desee engranar una ATN.
3. Programas de procesamiento y control:

IBIS :generación y lectura de los registros para las tablas
ENTE :lectura de los datos de entrada
SEGMENT :segmentación del texto de entrada
KAKADU :comprobación de si la palabra es funcional o no
SPATZ :realización de transformaciones (lápices-->lápiz,...)
WACHTEL :Es el programa de descomposición morfológica y consta de las funciones: MORPH (busca arborea),DELTA (acepta o rechaza las transiciones)y RUECKSETZ

Hay varios posibles outputs de la descomposición morfológica, interno -a engranar con la lematización-, u otros según el proceso a continuar.
Ya que el mecanismo ofrece varias posibles descomposiciones para algunas palabras (por razones diacrónicas o fonéticas), éstas son alistadas según un rango de validez estadística (ZSORT), y luego lingüística (ZERRAUS) por la que se eligen las descomposciones definitivas, si son varias en el caso de los homografos.
La lematización y categorización es un "matching" de los tipos y subtipos de los patrones morfológicos que tienen asignados cada morfema.

Diagrama del funcionamiento de los programas para la descomposicion morfologica y lematizacion.

Este paquete de programas esta implementado en SPL y ocupa unos 90 Kb(procesamiento y listas).
Actualmente existe una versión interactiva en un ordenador SIEMENS 7760.