Reconocedor automático de formas verbales que trata conjugación y pronombres enclíticos

Autores: O. Santana, Z. Hernández, G. Rodríguez, J. Pérez, F. Carreras, S. Bogliani.
Departamento de Informática y Sistemas
Universidad de las Palmas de Gran Canaria

Resumen

El propósito del presente trabajo consiste en describir una herramienta que reconoce automáticamente las diferentes formas conjugadas de un verbo, identificando su infinitivo, tiempo, número y persona –incluyendo las modificaciones dadas por la presencia de pronombres enclíticos.

0.– Introducción

Cada vez se hace más patente la necesidad de acercar la comunicación entre hombre y máquina al lenguaje natural. Se plantean en esta línea problemas de consultas a bases de datos documentales, generación y análisis de textos escritos, problemas de corrección ortográfica, etc. El análisis morfológico constituye una base sólida sobre la que es posible aproximarse con mayor rigor al sintáctico e incluso al semántico.

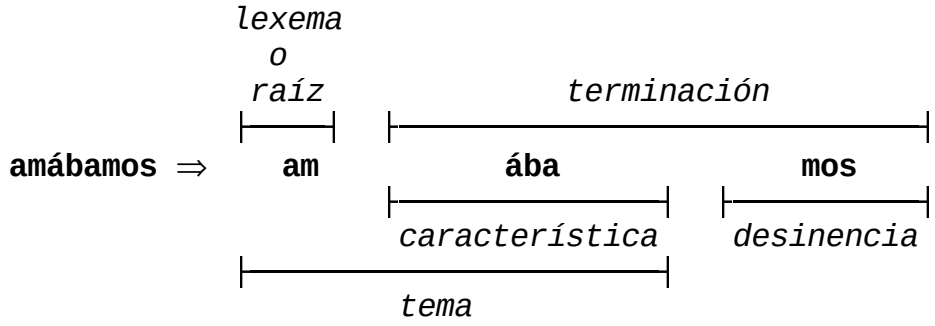
Se ha desarrollado un instrumento con la capacidad de determinar la raíz de una forma verbal dada e identificar su infinitivo y la conjugación que la afecta. El trabajo se enmarca dentro de un proyecto orientado a servir de ayuda en la elaboración de documentos escritos; básicamente consiste en una serie de herramientas dedicadas a proporcionar análisis del texto desde diferentes enfoques, y una de estas ópticas corresponde al estudio del empleo de las formas verbales.

En la sección 1 se hace una breve exposición sobre la morfología del verbo con la atención puesta en el tratamiento automático; los grupos de tiempos irregulares se tratan en la sección 2 también con la misma intención; en la sección 3 se exponen las características de los pronombres enclíticos; en la sección 4 se plantea de forma detallada el problema, concretándose su realización en la sección 5; la sección 6 presenta las conclusiones y perspectivas.

1.– Morfología del verbo

Por sus características formales, el verbo es aquella parte de la oración que contiene *morfemas* flexivos de número, persona, tiempo y modo. Suele aplicarse la denominación de *desinencias* a los morfemas de número y persona, y la de *características* a los de modo y tiempo. Si se suprime de una forma verbal su desinencia y característica, queda el *lexema* o *raíz*. La unión

de la raíz con la característica se denomina *tema*. Cualquier forma verbal se construye básicamente con una raíz y una *terminación* –constituida por los morfemas.



Para el tratamiento automatizado de la conjugación no conviene olvidar que la mayor parte de los verbos son regulares y mantienen la raíz invariable en su conjugación –si se exceptúa la posición del *acento de intensidad*. Los verbos irregulares presentan variaciones en su raíz –vocálicas, consonánticas o mixtas– además de otros tipos de anomalías.

– La IRREGULARIDAD VOCÁLICA consiste en el cambio de una vocal por otra u otras:

e por **i** **concebir** ⇒ **concibo**

o por **ue** **colgar** ⇒ **cuelgo**

– La IRREGULARIDAD CONSONÁNTICA tiene lugar si se reemplaza una consonante por otra:

c por **qu** delante de **e** **caber** ⇒ **quepo**

o se añade una consonante a la consonante final de la raíz del infinitivo:

añadir **z** **conocer** ⇒ **conozco**

– La IRREGULARIDAD MIXTA aparece cuando se sustituyen una vocal y una consonante por otra vocal y otra consonante:

ab por **ep** **saber** ⇒ **sepa**

o en la adición del grupo **ig** a la última vocal de la raíz:

añadir **ig** **caer** ⇒ **caigo**

Existen irregularidades que afectan al tema –**di** de **decir**, **haz** de **hacer**, **doy** de **dar**, **voy** de **ir**, etc.– y otras de más difícil sistematización:

– contracciones: **ver** ⇒ **ves** en vez de **vees**.

– verbos con más de una raíz: **ser** ⇒ **fueron**, **era**.

– participios y gerundios irregulares: **escribir** ⇒ **escrito**, **dormir** ⇒ **durmiendo**

No se consideran propiamente irregularidades los simples cambios ortográficos:

– **g** por **gu** delante de **e** o **i** **sigo** ⇒ **sigue**

– **qu** por **c** delante de **a** u **o** **delinquir** ⇒ **delinca**

En estos casos no ha variado el *fonema*, sino el carácter que lo representa. Otras aparentes anomalías obedecen a principios generales del sistema fonológico español y tampoco se consideran irregularidades **–leyendo** de **leer**.

2.– Grupos de tiempos irregulares

La irregularidad afecta siempre a más de un tiempo. Son tres los grupos de tiempos que comparten la misma irregularidad:

– Si es irregular el presente de indicativo, también lo son los otros presentes.

– Si es irregular el pretérito indefinido, poseen la misma irregularidad el pretérito imperfecto de subjuntivo y el futuro imperfecto de subjuntivo.

– Si es irregular el futuro imperfecto de indicativo, tiene la misma irregularidad el condicional simple.

El pretérito imperfecto de indicativo carece de irregularidades, salvo raras excepciones que se reducen casi exclusivamente a los imperfectos heredados del latín **era** e **iba**, de los verbos de *raíz múltiple* **ser** e **ir**.

Los tiempos compuestos y la voz pasiva no presentan más irregularidades que las de los auxiliares **haber** y **ser** y las de algunos participios como **escrito**, **impreso**,...

3.– Pronombres enclíticos

El empleo de los pronombres enclíticos ha ido reduciéndose con el tiempo, y ahora se usan casi exclusivamente cuando el verbo está en infinitivo, imperativo o gerundio.

Un verbo puede llevar simultáneamente hasta tres pronombres átonos; en tal caso, la partícula **se** debe preceder al resto, el de segunda persona **–te**, **os–** adelanta siempre al de primera **–me**, **nos–** y cualquiera de estos dos antecede al de tercera **–le**, **les**, **la**, **las**, **lo**, **los**.

Al unirse los pronombres enclíticos con el verbo, se producen algunas alteraciones que afectan al último carácter del verbo:

– delante del enclítico **nos** se pierde la **s** de la primera persona del plural del subjuntivo e imperativo

comamos + **nos** ⇒ **comámonos**

esta pérdida se produce también en otros tiempos del verbo, pero el pronombre enclítico es muy poco usual en ellos. En la segunda persona del plural seguida del enclítico **se** desaparece una **s**

comamos + **selo** ⇒ **comámoselo**

– la forma reflexiva de la segunda persona del plural del imperativo pierde la **d**

comed + **os** ⇒ **comeos**

3.1.– El acento ortográfico

El acento ortográfico de las formaciones con enclíticos está siempre de acuerdo con las reglas generales cuando dicha formación es esdrújula o sobresdrújula. Cuando una forma verbal llana o esdrújula se agrupa con uno o más enclíticos, la vocal prosódicamente acentuada del verbo lleva siempre tilde, lo exija o no cuando se emplea sin enclíticos: **decía-me-lo**, **veía-la**, **oía-lo –**de acuerdo con las formas verbales **decía**, **veía**, **oía–**, **hablába-se**, **mirándo-os**, **quisiéra-lo –**en contraste con **hablaba**, **mirando**, **quisiera**. Y cuando una forma verbal aguda –incluyendo las monosilábicas– se agrupa con dos enclíticos, la vocal prosódicamente acentuada del verbo se escribe siempre con tilde, lo requiera o no cuando se emplea sola: **partió-se-le**, **oír-se-lo**, **dé-se-la –**de acuerdo con **partió**, **oír**, **dé–**, **dá-se-lo**, **dí-me-lo**, **decíd-nos-lo**, **pedír-me-la –**en contraste con **da**, **di**, **decid**, **pedir**.

El acento ortográfico de las formaciones con enclíticos deja de estar de acuerdo con las reglas generales en algunos casos en que la formación resulta con acentuación llana. En las formas verbales agudas –incluyendo las monosilábicas– seguidas de un sólo enclítico: **da-le**, **fui-me**, **decid-me**, **reír-se, oír-lo**; **dé-le**, **salí-me**, **partió-se** el verbo conserva su acento ortográfico originario; solamente en los cinco primeros ejemplos la formación se atiene a las reglas generales del uso ortográfico (**dale** como **sale**, **fuime** como **fuiste**,...) y en los tres últimos ejemplos diverge de las reglas generales, ya que una palabra llana terminada en vocal normalmente no lleva tilde.

Aparecen también sometidos a un régimen ortográfico especial los imperativos plurales de los verbos reflexivos, o en construcción reflexiva, tras la pérdida de la desinencia **d**. Formas como **marcha-os**, **detene-os** han de emplearse sin tilde, a pesar de que la forma verbal es aguda y de que se agrupa con sólo un enclítico. Sin embargo los verbos de la tercera conjugación llevan tilde a causa del hiato como **partí-os**, salvo la forma **id-os** del verbo **ir** en el que **í-os** es un arcaísmo.

4.– Del problema

Encontrar de una manera automática el infinitivo correspondiente a una forma verbal dada es una tarea simple cuando se trata con verbos regulares; en caso de los irregulares la cosa resulta algo más compleja. Una solución podría ser la de construir una estructura que contuviera conjugados todos los tiempos simples en todas las personas y números; sin embargo, tal situación conllevaría el mantenimiento de una organización excesivamente grande si se tiene presente que cada infinitivo no defectivo da lugar a sesenta y dos formas verbales o más –aquellos que presentan múltiples formas para una misma persona.

Un primer estudio acerca de la morfología de los verbos irregulares pone de manifiesto que son excepcionales **–ir**, **ser–** los que usan un número de cambios de raíces superior a los tres que se corresponden con los grupos de tiempos reseñados en la sección 2 y que la mayoría no presenta irregularidades en todos los grupos

caer añade **ig** ⇒ **caigo** sólo en el primer grupo

e incluso algunos repiten el mismo modelo de cambio para más de uno

concebir ⇒ **concibo**, **concibiera**

en el primer y segundo grupo cambia **e** por **i** según una irregularidad vocálica.

Como por un lado el número de raíces diferentes de cada verbo irregular es bastante menor que el de sus formas verbales y por otro las terminaciones posibles son limitadas y sabidas, resulta factible afrontar la solución en dos fases: 1) por criterios puramente estructurales separar de la forma verbal raíz y terminación y 2) localizar la raíz en un conjunto de raíces irregulares conocidas y, en caso de encontrarla, relacionarla con la raíz de su infinitivo.

Debe entenderse en su justa medida la expresión de que el corte raíz/terminación se realiza por criterios puramente estructurales, porque la solución, con la intención de obtener un número óptimo de raíces, pasa por elegir como raíz la que resulte de tomar la terminación más larga que se pueda encontrar, siempre que no sea posible elegir la del infinitivo o una anteriormente tomada para ese verbo; ello implica que en algunos casos la raíz extraída no se corresponda con la lingüística.

Ante formas que coinciden completamente con una terminación **–**la terminación **-ase** del pretérito imperfecto de subjuntivo coincide con la forma **ase** del verbo **asar–** se toma como raíz parte del principio de la terminación; el extender esta norma al caso de raíces que quedan reducidas a una sola letra **v**-**oy**, ha contribuido a eliminar ciertas ambigüedades, por ejemplo con **v**-**er**.

Por último se ha tenido en cuenta: a) que algunas formas verbales pueden pertenecer a más de un infinitivo

fueron es una forma verbal de **ir**, pero también de **ser**

y b) reconstruir el infinitivo acentuado. Los infinitivos son siempre agudos y no se acentúan salvo en la tercera conjugación cuando la raíz gramatical acaba en vocal fuerte **–oír**, **reír–** para indicar el hiato.

4.1.– De los pronombres enclíticos

Las formas verbales afectadas por pronombres enclíticos necesitan deshacerse de tales accidentes para poder tratar su forma no pronominalizada.

Si la terminación coincide con alguno de los pronombres o combinaciones de pronombres considerados, se elimina dicha terminación y se retoca la forma resultante en el caso de que pudiera estar afectada por alguna de las alteraciones consideradas en la sección 3.

Teniendo en cuenta que la sílaba tónica debe mantenerse, se estudia para todas las formas obtenidas la conservación o desaparición de la tilde, y se busca como forma verbal conjugada que admita pronombres enclíticos.

El procedimiento que se encarga del estudio de la tilde emplea un separador silábico y un comprobador de las reglas ortográficas de acentuación que han sido desarrollados al efecto; permite realizar la separación en sílabas de una palabra, determinar cuál es su sílaba tónica y cambiar de posición, situar o hacer desaparecer la tilde dada la sílaba tónica.

5.– Realización

El proceso de identificación se organiza en dos fases: primero se reconocen los posibles pronombres enclíticos y luego la forma o formas verbales resultantes; el resultado será el infinitivo (o posibles infinitivos) acompañado de los tiempos, números y personas (y en su caso de la combinación de pronombres enclíticos) que dan lugar a la forma original.

La primera fase consiste en tratar la existencia de los posibles pronombres enclíticos y generar la forma o formas resultantes de su eliminación considerando las posibles alteraciones. Para ello las posibles combinaciones de pronombres convenientemente codificadas se organizan en una estructura.

La ejecución de la segunda fase hace uso de una estructura que mantiene una relación de todas las terminaciones verbales adecuadamente clasificadas. Se determinan las posibles terminaciones para una forma dada. Se obtienen los infinitivos al disponer de todas las raíces de los verbos en una organización que permite realizar sus búsquedas exactas de forma eficiente; se comprueba con facilidad si la terminación dada le corresponde y en tal caso cual es la flexión que representa.

La aplicación se ha desarrollado en C++ bajo Windows. Las estructuras de datos utilizadas por el programa ocupan un total de 853Kb en disco. Sobre un procesador Intel 486 a 50Mhz con 8Mb de memoria RAM, se identifican en 96,6 minutos las 732.609 formas conjugadas derivadas de 11.830 infinitivos; si se elimina el tiempo de lectura del fichero de datos –de 10,3 Mb– y el de presentación de resultados –2,1 minutos en total– resulta una velocidad de reconocimiento de 129,2 formas por segundo. Aunque entre las formas verbales generadas no se han introducido pronombres enclíticos, el reconocedor incluye la posibilidad de su existencia; de no ser así, la velocidad de reconocimiento sería 170,1 formas por segundo. Al aplicar este proceso a un texto de carácter literario con 111.691 palabras la velocidad con la que se descartan las palabras que no son formas conjugadas alcanza las 260,3 por segundo y el texto se procesa a 169,7 palabras por segundo.

6.– Conclusiones

Se ha desarrollado una herramienta que permite identificar el tiempo, la persona, el número y los pronombres enclíticos que pueda poseer una forma verbal conjugada cualquiera. Próximamente se incluirán las flexiones correspondientes a la sustantivación, adjetivación y adverbialización de formas verbales. En una etapa posterior, se dispondrá de un generador de las formas verbales que a partir de un infinitivo y una flexión (conjugación con sus pronombres enclíticos, sustantivación, adjetivación, adverbialización,...) sea capaz de obtener la forma correspondiente.

Es factible la integración del reconocedor en desarrollos relacionados con la flexión de palabras en general y de los verbos en particular, permitiendo análisis de textos de una gran profundidad y nivel de detalle. Resulta de interés su inclusión en la realización de búsquedas complejas en las que se desee actuar con independencia de la flexión que afecte a la forma considerada.

Agradecimientos

Queremos agradecer al profesor Dr. Manuel Alvar Ezquerra del Departamento de Filología Española I de la Universidad de Málaga su colaboración en cuantas consultas le hemos formulado a lo largo del desarrollo del presente trabajo.

Referencias

ALSINA, R. (1990): *Todos los Verbos Castellanos Conjugados*, Teide, 17ª Edición.

ALVAR EZQUERRA, M. (1993): *La formación de palabras en español*, Ed.: Arco/Libros, Madrid.

VOX (1990): *Diccionario Actual de la Lengua Española*, Ed.: Biblograf S.A., Barcelona.

GILI GAYA, S. (1985): *Curso superior de sintaxis española*, Vox, Biblograf, Barcelona.

GÓMEZ TORREGO, L. (1991): *Manual de Español Correcto*, Arco/Libros S.A., Madrid.

GÓMEZ TORREGO, L. (1992): *El buen uso de las palabras*, Arco/Libros S.A., Madrid.

LYONS, J. (1975): *Nuevos horizontes de la lingüística*, Alianza Editorial, Madrid.

MARTINET, A. (1978): *Elementos de lingüística general*, Gredos, Madrid.

REAL ACADEMIA ESPAÑOLA (1989): *Esbozo de una nueva gramática de la lengua española*, Espasa Calpe, Madrid.

G. RODRÍGUEZ, Z. HERNÁNDEZ, O. SANTANA, «Agrupaciones de Tiempos Verbales en un Texto», en *Anales de las II Jornadas de Ingeniería de Sistemas Informáticos y de Computación*, Quito (Ecuador), Abril 1993, págs. 132-137.

O. SANTANA, Z. HERNÁNDEZ, G. RODRÍGUEZ, «Conjugaciones Verbales», en *Boletín de la SEPLN*, Nº 13, Febrero 1993, págs. 443-450.

O. SANTANA, J. C. RODRÍGUEZ, J. D. GONZÁLEZ, «FRECTEXT: Una Aplicación de Ayuda a la Elaboración de Documentos», *Boletín de la SEPLN*, Nº 13, Febrero 1993, págs. 451-462.

SECO, M. (1991): *Diccionario de dudas y dificultades de la lengua española*, Espasa Calpe, 9ª Edición, Madrid.