Diccionarios del castellano en ordenador para la composición y verificación de textos

L.Sopeña
Centro de Investigación UAM-IBM
Paseo de la Castellana 4
28046 Madrid

Introducción

El procesamiento de documentos por ordenador dispone hoy en día de numerosas herramientas que permiten al usuario introducir un texto en memoria, darle formato e incluso, para algunos idiomas (en particular el inglés), verificar la ortografía, la puntuación y el estilo **[1, 2, 3, 4]**. En cambio, otras lenguas (el castellano entre ellas), han recibido mucha menor atención a este respecto **[5]**.

El presente artículo describe brevemente un trabajo en desarrollo cuyo objetivo es construir un entorno integrado para la composición y verificación de documentos escritos en castellano **[6]**, para lo cual no existen en estos momentos programas similares disponibles.

En una primera fase se ha completado la elaboración de un diccionario del castellano. Esta es una tarea de interés múltiple, ya que es imprescindible disponer de un diccionario en cualquier aplicación en que intervenga el Lenguaje Natural. Actualmente, el diccionario se encuentra terminado en una versión que incluye unas 35.000 raíces que, flexionadas, dan lugar a más de 400.000 palabras distintas. Junto con este léxico básico se ha construido un diccionario de sinónimos que cuenta con más de 15.000 entradas.

En las próximas secciones se resumen las características de ambos diccionarios y la forma en que son utilizados por un programa de composición y verificación de textos. Se describen también las ampliaciones previstas para el mismo, entre las que se incluye el desarrollo de un analizador sintáctico del castellano.

El diccionario de formas flexionadas

Para decidir en una primera fase los términos que debían incluirse en el diccionario se partió de un análisis de frecuencia llevado a cabo sobre diferentes textos previamente seleccionados y grabados, artículos de prensa, novelas, ensayos, etc. hasta un total aproximado de un millón de palabras. Se estudió también el listado completo de las entradas del **Diccionario de la Real Academia Española [7] (DRAE)**, así como numerosos otros diccionarios impresos **[8, 9, 10, 11]**. La información así obtenida se clasificó y filtró, teniendo en cuenta el primer objetivo a que se iba a destinar el diccionario; el corpus debía cubrir el *lenguaje escrito usual*, y en este ámbito debía dar cuenta del mayor número posible de palabras.

El diccionario consiste en una lista de palabras flexionadas, sin definiciones asociadas, cada una de las cuales dispone de un conjunto de informaciones adicionales: género, número, tiempo, persona, modo, etc. **[12]**. En general, las palabras que pertenecen a dominios técnicos o especializados (medicina, derecho, ingeniería, lingüística, etc.) no han sido incluidas. Tampoco lo han sido los términos exclusivamente coloquiales o de argot, ni los usos regionales del castellano (como el voseo argentino: *tenés, querés*), o muy específicos (como el futuro de subjuntivo: *tuviere, quisiere*, reservado hoy a escritos legales). Se han excluido también muchas formas derivadas, tales como diminutivos, despectivos, superlativos (no así las formas irregulares); en cuanto a los adverbios en **-mente**, sólo se han listado los de uso más frecuente.

Dentro de lo que podemos llamar inventarios abiertos se han incluido los adjetivos calificativos, nombres comunes y verbos más usuales en lenguaje escrito. Los inventarios cerrados, por su parte, comprenden las partes fijas de la oración. También se han incluido los nombres propios de persona y apellidos españoles más frecuentes, y se ha puesto especial interés en que la información relacionada con la geografía, tanto física como política, fuera lo más completa posible en lo que se refiere a los países de habla hispana.

Figuran los nombres de todos los países del mundo y sus capitales, así como las ciudades extranjeras cuyo nombre castellano difiere del original, como Milán, Burdeos, Munich, etc. También han sido considerados los neologismos y latinismos más generalizados en castellano. En el caso de los extranjerismos, si la palabra tiene mucha aceptación, aparece, además de la grafía original, la grafía adaptada a las normas ortográficas del español.

La información sobre el léxico está contenida en dos ficheros principales: el de formas base y el de morfemas flexivos, que se describen en las secciones siguientes.

Fichero de formas base

Incluye la lista completa de las formas que se acaban de describir, con especificación de la forma base en que realizan sus flexiones. El uso de apuntadores permite referencias al fichero de morfemas flexivos.

Cada entrada dispone de las especificaciones siguientes:

1. Categoría funcional, es decir, verbo, nombre, adjetivo, adverbio, preposición, conjunción, artículo, pronombre, interjección; las palabras con más de una parte de la oración asociada tendrán tantas marcas como categorías.

2. En los verbos, muy complejos debido al elevado número de irregularidades que presentan y lo difícil de su clasificación, se señala si son transitivos, intransitivos o auxiliares. También se han incluido las combinaciones más frecuentes de verbos y pronombres enclíticos [13]. Por otro lado, se ha previsto la posibilidad de añadir códigos para caracterizar su comportamiento y utilización al nivel superficial: complementos, preposiciones, etc.

3. Hay marcas adicionales para describir otras características de cada término, tales como si la palabra está incluida o no en el DRAE, si se trata de un término extranjero, de una palabra latina, un nombre propio, geográfico, etc.

Fichero de morfemas flexivos

Especifica los morfemas derivativos empleados en la generación de las formas flexionadas a partir de las entradas del fichero de formas base; se ha definido una lista de paradigmas para cada categoría de nombres, adjetivos y verbos, para dar cuenta de cada uno de los modelos de flexión.

La clasificación tiene en cuenta los problemas derivados del tratamiento automático de las flexiones, es decir que considera como irregularidades algunos comportamientos no tratados tradicionalmente como tales. Por ejemplo, casos puramente fonéticos, como z → c ante e, i (p.e. *cazar* → *cace*), y casos relacionados con los signos diacríticos, tanto la diéresis (p.e. *avergonzar* → *avergüenzo*), como los acentos (p.e. *joven* → *jóvenes, carácter* → *caracteres*).

Además es necesario considerar casos de flexiones incompletas (p.e. el adjetivo *alisios* sólo existe en masculino plural; el nombre *afueras* sólo existe en femenino plural). En cuanto a los verbos, un tipo paralelo de irregularidad se presenta en los llamados defectivos (p.e. *llover, abolir, pudrir, etc.*). Finalmente, hay palabras con más de una realización en una de sus formas; ello ocurre con algunos nombres (p.e.

varíz/varice, ambas correctas en femenino singular) y con todos los verbos en el imperfecto de subjuntivo (p.e. *saliera/saliese*) y en algunas otras formas aisladas (p.e. el imperativo *satisfaz/satisface*); asimismo, algunos adjetivos presentan un fenómeno similar de formas dobles según se encuentren antes o después del nombre (p.e. *buen/bueno, mal/malo*).

Junto a los adjetivos con marca de género y número (p.e. *rojo, roja*), hay otros sin ellas (p.e. *amable*), cuyo género se define según el del nombre al que modifican. Entre ellos, algunos funcionan en contextos fijos y restringidos, y se definen porque modifican a nombres masculinos o femeninos (p.e. *ácimo*, asociado únicamente a *pan*).

Debe notarse que el gran número de irregularidades en el mecanismo de flexión ha obligado a detallar cada una de ellas, ya que no podían incluirse en ninguno de los modelos generales; por esta razón, numerosos paradigmas incluyen un número muy pequeño de casos, y a veces, incluso, uno solo.

El diccionario de sinónimos

Para construir el diccionario de sinónimos se ha utilizado un volumen publicado [14], que ha tenido que ser modificado debido a las necesidades específicas del tratamiento informático y de los numerosos errores tipográficos y de coherencia encontrados en su contenido. Ello ha permitido estudiar de modo completo la sinonimia y llevar a cabo un estudio exhaustivo de uno de los diccionarios de sinónimos del castellano más conocidos.

En primer lugar se ha mantenido la congruencia de este léxico con el previamente descrito, de modo que todo término del diccionario de sinónimos esté también incluido en el diccionario base.

La necesidad de mantener la coherencia semántica en el interior del diccionario de sinónimos ha sido uno de los primeros objetivos, ha mostrado el poco rigor con que se construyen los diccionarios impresos y ha permitido la aplicación de pruebas sistemáticas y modificaciones en nuestra versión con el objeto de conservar la simetría, tener en cuenta la hiperonimia y mantener las referencias cruzadas dentro de unos límites semánticamente válidos.

A partir de las marcas sintácticas del diccionario de formas flexionadas, una entrada del diccionario de sinónimos aparecerá tantas veces como partes de la oración se le hayan asignado. Por ejemplo:

circular: j
redondo, curvo, curvado.
circular: nf
orden, aviso*, notificación, carta, nota.
circular: v
andar, moverse, transitar*, pasear, deambular;
divulgarse, propagarse, expandirse, difundirse.

Además, dentro de una misma parte de la oración, los sinónimos se agrupan según las distintas acepciones de la palabra; también se permiten las referencias cruzadas (marcadas con un asterisco * en el fichero), que encadenan un sinónimo con otra entrada del diccionario, extendiendo así el poder de información del léxico.

Puede definirse también información más específica sobre las entradas por medio de lo que hemos llamado cualificadores, que introducen restricciones adicionales en la entrada para que se aplique tal acepción. Ejemplo:

```
costa: N
        playa, litoral, margen, orilla, borde;
        < plural >
          cargas, desembolso, importe.
echar: V
        expulsar, repeler, rechazar, despachar, excluir;
        deponer, destituir;
        dar, entregar, repartir;
        .....
        < se >
          tenderse, acostarse, tumbarse, arrellanarse.
```

Composición de textos basada en los diccionarios

Los diccionarios que se acaban de describir son utilizados por un programa de tratamiento de textos con el objeto de asistir al usuario en la composición de sus documentos. A continuación se presentan las principales funciones que los diccionarios permiten.

Verificación ortográfica

El método se basa en la identificación de todas las cadenas del texto que no se encuentren en el diccionario. Los algoritmos de verificación aíslan cada palabra (*'token'*), la buscan en el léxico e indican al usuario las que no están en él (resaltándolas en la pantalla o empleando un color diferente para ellas). Un *'token'* es por tanto toda secuencia de letras enmarcada por delimitadores (blanco, coma, punto, punto y coma, dos puntos, guión, abrir y cerrar signos de exclamación e interrogación). Así, el tamaño del diccionario tendrá varias implicaciones obvias: la frecuencia de palabras correctas rechazadas, el tiempo de búsqueda, la cantidad de memoria necesaria. Un compromiso entre todos estos factores y el uso de diversos mecanismos de compactación permiten que el tamaño de los diccionarios permanezca dentro de unos límites razonables, a la vez que proporcionan una cobertura muy considerable del léxico.

La verificación ortográfica que se lleva a cabo en este momento considera cada palabra del texto de forma independiente a las demás, sin tener en cuenta su función concreta dentro de la oración.

Una posibilidad adicional e interesante del programa es que permite al usuario definir su propio diccionario de addendas, en el que puede almacenar los términos que emplee habitualmente y sean desconocidos por el sistema (nombres propios, palabras técnicas o específicas). Una vez definido este léxico adicional, sus entradas son incorporadas al vocabulario conocido por el sistema y serán, en lo sucesivo, consideradas correctas.

Corrección ortográfica

Además de detectar las palabras incorrectas del texto, el sistema también puede proponer para cada *'token'* erróneo del texto, previa solicitud del usuario, una lista de candidatos, palabras muy similares a él que sí se encuentran en el diccionario. La lista se presenta en una ventana junto a la cadena incorrecta con los términos alternativos ordenados en orden decreciente de prioridad, según el valor de un cierto índice de similitud que se calcula para cada palabra. La "similitud" entre dos palabras se determina algorítmicamente, y depende fundamentalmente del número de alteraciones a que hay que someter al *'token'* para obtener la palabra correcta. Por tanto es función de factores como la diferencia de longitudes entre ambas cadenas, la diferencia en las secuencias respectivas de caracteres (debida a alguno de los errores mecanográficos típicos: transcripción, omisión, inserción y sustitución de letras), la coincidencia de la última letra, etc. El usuario puede escoger una de las palabras de la lista propuesta y el sistema la pondrá automáticamente en lugar del término erróneo.

Función de morfología

Para cada palabra del texto el programa puede encontrar todas sus posibles formas base con su parte de la oración respectiva (actualmente, fuera de contexto). Puede también generar la colección completa de formas derivadas para cada una de tales posibilidades. En castellano esta función es especialmente útil en el caso de flexiones poco usuales, como verbos irregulares o defectivos, cuando se tienen dudas sobre el uso del acento, con algunos nombres y adjetivos especiales, con términos de escasa utilización, etc.

Función de sinónimos

El mecanismo es muy similar al que acabamos de describir para los términos alternativos: cuando el usuario pide sinónimos para una palabra del texto, éstos le son presentados en una ventana junto a aquélla. Actualmente, las palabras que tiene más de una parte de la oración aparecen en la pantalla con los sinónimos correspondientes a cada una de ellas, con independencia de la función gramatical que esa palabra tenga en el contexto concreto en que se esté empleando. Por ejemplo, los sinónimos para *bajo* se presentarán en varias listas: como verbo, como nombre, como adjetivo, como adverbio y como preposición.

El usuario podrá escoger uno de los sinónimos y colocarlo en el texto automáticamente en sustitución de la palabra que figuraba inicialmente. En la fase actual, el sistema no flexiona los candidatos a la forma del *'token'* original. A partir de él, lo que hace es llevar a cabo un análisis morfológico, encontrar su raíz y buscar los sinónimos que le corresponden en el diccionario. Así, si el usuario escribe la frase *Juan quiere a María* y pide sinónimos para *quiere*, el sistema será capaz de deducir la forma base *querer*, consultará con

ella el diccionario de sinónimos y mostrará, por ejemplo, el infinitivo *amar*, pero no *ama*, que es la forma conjugada correspondiente al verbo original.

Análisis sintáctico y otras ampliaciones

La composición de textos basada en diccionarios que acabamos de describir supone una gran ayuda para la escritura de documentos, pero es claramente insuficiente. Nuestra siguiente etapa es la construcción de una analizador sintáctico del castellano que se integrará, como primera aplicación, en el sistema descrito. Ello tendrá varias implicaciones en cuanto a la ampliación de sus posibilidades y le añadirá nuevas capacidades de verificación.

Por ejemplo, va a permitir el tratamiento de expresiones idiomáticas y frases hechas, palabras compuestas, expresiones adverbiales, etc. Va a hacer posible que la función de sinónimos proponga los sinónimos para una palabra solamente en la parte de la oración adecuada, excluyendo las demás posibilidades de acuerdo con el contexto.

Se van a poder también superar las limitaciones de la verificación ortográfica actual, teniendo en cuenta el contexto; así, los errores debidos al uso de palabras correctas (es decir, contenidas en el diccionario) en un entorno sintáctico incorrecto se podrán detectar en la mayoría de los casos. El mayor número de fuentes de confusión que hoy pasan desapercibidas y que se van a poder detectar son debidas fundamentalmente a tres tipos de ambigüedad:

- Ambigüedad gráfica: palabras homófonas con una diferencia gráfica en el acento y con distintas partes de la oración (*este/éste, cuanto/cuánto, de/dé, si/sí*).
- Ambigüedades de acentuación: basadas en el cambio de acento dentro de un grupo de palabras, a veces con cambio asociado de función gramatical (*baile/bailé, frío/frió, cántara/cantara/cantará, ame/amé*).
- Ambigüedades fonéticas: debidas de los problemas ortográficos derivados de la fonética castellana (*asta/hasta, tubo/tuvo, callado/cayado, contexto/contesto*).

Naturalmente éstas serán solamente algunas de las aplicaciones más inmediatas del analizador, si bien algunas de las ambigüedades que se acaban de describir necesitarán además de conocimientos semánticos para poder ser resueltas; esta posibilidad adicional no la estamos considerando por el momento. Otros usos obvios incluyen la detección de errores de concordancia dentro de un sintagma nominal, entre el sujeto y el verbo de una oración, errores en el uso de pronombres como leísmo y laísmo, errores en el orden de los pronombres, etc.

Los diversos elementos que integran el sistema que acaba de ser descrito constituyen un conjunto de piezas diferentes cuya aplicación no está limitada, por supuesto, a la composición de documentos. Se han previsto otros objetivos para los diccionarios y el analizador; por ejemplo, ya se ha desarrollado un sistema de ayuda a la conjugación de verbos para estudiantes de gramática española. Otras ideas incluyen un sistema de elaboración automática de resúmenes, la inclusión en el diccionario de definiciones para cada término, así como su traducción a otros lenguajes, y finalmente la revisión del estilo de los documentos.

Referencias

[1] Larson, J. A.: *End User Facilities in the 1980's*, (Chapter 6: Creating, Revising, and Publishing Office Documents), IEEE, New York, 1982.

[2] Cherry, L.: Writing Tools, *IEEE Trans. on Communications*, vol. 30, no. 1, January 1982.

[3] Peterson, J.L.: Computer Programs for Detecting and Correcting Spelling Errors, *Comm. of the ACM*, December 1980, vol. 23, no. 12.

[4] Richardson, S.: Enhanced Text Critiquing using a Natural Language Parser, IBM Research, RC 11332, August 1985.

[5] Sopeña, L.: Lingüística y Procesamiento de Textos en Castellano, *PC World*, no. 8, febrero 1986.

[6] Sopeña, L.: Procesamiento de Textos en el Centro de Investigación UAM-IBM, *Procesamiento del Lenguaje Natural*, no. 3, mayo 1985.

[7] Real Academia Española: *Diccionario de la Lengua Española*, vigésima edición, Ed. Espasa Calpe, Madrid, 1984.

[8] Moliner, M.: *Diccionario de uso del español*, Ed. Gredos, Madrid, 1982.

[9] Casares, J.: *Diccionario ideológico de la Lengua Española*, Ed. Gustavo Gili, Barcelona, 1982.

[10] *Diccionario Anaya de la Lengua*, Ed. Anaya, Madrid 1980.

[11] Seco, M.: *Diccionario de dudas y dificultades de la lengua española*, Ed. Espasa Calpe, Madrid, 1986.

[12] Casajuana, R., Rodríguez, C.: Verificación ortográfica en castellano; la realización de un diccionario en ordenador, *Español Actual*, no. 44, 1985.

[13] Casajuana, R., Rodríguez, C.: Clasificación de los verbos castellanos para un diccionario en ordenador, *Actas 1er. Congreso de Lenguajes Naturales y Lenguajes Formales*, Barcelona, octubre 1985.

[14] Sainz de Robles, F. C.: *Diccionario español de sinónimos y antónimos*, Ed. Aguilar, 1984.