Corrección gramatical y Preprocesamiento

Flora Ramírez Bustamante
Fernando Sánchez León
Laboratorio de Lingüística Informática
Facultad de Filosofía y Letras
Universidad Autónoma de Madrid
E-28049 Madrid (Spain)
{flora;fernando}@maria.lllf.uam.es

Thierry Declerck
DFKI GmbH
(German Research Center for Artificial Intelligence)
Stuhlsatzenhausweg 3
D-66123 Saarbruecken (Germany)
declerck@dfki.uni-sb.de

Resumen

Este artículo explora los niveles previos al procesamiento dentro del ámbito de la verificación gramatical con el fin de reflexionar sobre la robustez de los analizadores sintácticos que utilizan técnicas de PLN. Sobre la experiencia adquirida durante la realización del proyecto GramCheck, se intenta una primera caracterización de errores susceptibles de un tratamiento a bajo nivel, es decir, en fases previas al análisis profundo que proporcionan plataformas como ALEP.

Palabras clave: Corrección gramatical, preprocesamiento lingüístico, *parsing* robusto

1 Motivación

Durante la ejecución del proyecto LS-GRAM[1], se dio gran importancia al componente de preprocesamiento (*Text Handling, TH*) de la plataforma ALEP. Haciendo uso de este subsistema, se han llevado a cabo algunos experimentos sobre preprocesamiento de cadenas de caracteres. Estos experimentos han sido positivos, pues han supuesto una ampliación de la cobertura y una mejora de la eficiencia de las gramáticas de LS-GRAM (véase [Bredenkamp *et al.* (1996)] y [Schmidt et al. (1996)]). Para la gramática del alemán, se han llevado a cabo algunos experimentos adicionales sobre la integración de la información proporcionada por `mpro`, una herramienta de análisis morfológico y un etiquetador (véase [Maas (1996)] y [Maas y Hirschfeld (1996)]). Los resultados positivos de este trabajo pueden consultarse en [Declerck y Maas (1997)].

En el contexto de esta estrategia de trabajo, surge la cuestión sobre qué aspectos lingüísticos pueden tratarse eficientemente durante el preprocesamiento en sistemas de PLN (no solo en lo que concierne a la plataforma ALEP). Puesto que los experimentos realizados con el componente TH no estaban encaminados especialmente a mejorar la robustez del sistema, este artículo se centra precisamente en el modo en que este aspecto podría tratarse también dentro del paradigma antes mencionado.

El campo de la *verificación gramatical* parece propicio para tales investigaciones, no solamente por su relación directa con la robustez necesaria de los sistemas —lo ideal sería mantener el analizador lingüístico libre de información incorrecta (quizá un deseo inalcanzable)—, sino también porque existe un proyecto, GramCheck[2], que ha utilizado ALEP como plataforma de

1.8em [1]El proyecto LS-GRAM (Large-Scale GRAMmars for EC languages) fue financiado por la CEC, LRE 61029.
1.8em [2]El proyecto GramCheck fue financiado por la CEC, MLAP93-11.

desarrollo de un corrector gramatical. Este proyecto definía un conjunto de procedimientos de corrección gramatical de alto nivel que requería una descripción también de alto nivel de las estructuras gramaticales. Por tanto, la cuestión que se plantea ahora es si es posible formular algunos de estos procedimientos en el nivel de análisis del preprocesamiento[3].

Este artículo discute estos aspectos, y no pretende ofrecer soluciones mágicas sino contribuir a la discusión, necesaria, sobre robustez y corrección gramatical en sistemas de PLN.

En la sección 2, se revisa la situación de la verificación gramatical con el fin de situar la estrategia anterior y la actual. GramCheck se presenta, pues, en la sección 2.1. También en esta sección se describen algunos de los problemas encontrados durante la investigación realizada. Los verificadores gramaticales tratan, entre otras cuestiones, errores de concordancia. Pero, ¿es posible detectarlos a un nivel bajo del procesamiento? Esto parece, al menos, problemático para el español. En la sección 3 presentamos una lista de fenómenos incorrectos en español, e intentaremos identificar posibles candidatos para la verificación en el nivel del preprocesamiento (algunos de estos fenómenos ya han sido tratados en un nivel más alto de descripción dentro de GramCheck).

2 Estado actual de la verificación gramatical

Los sistemas que tratan desviaciones gramaticales se centran principalmente en la integración de técnicas de procesamiento especiales para detectar y corregir, cuando es posible, tales desviaciones. Estas técnicas se incorporan, habitualmente, a las estrategias tradicionales de procesamiento, como en el caso de la relajación de rasgos en formalismos basados en unificación ([Bolioli *et al.* (1992)], [Ramírez y Sánchez-León (1996b)]), o la adición de un conjunto de reglas especiales que manejan construcciones erróneas ([Heidorn *et al.* (1982)], [Thurmair (1990)], [TWB (1992)]). Estos sistemas presentan una aproximación lingüísticamente motivada (mediante la formulación de reglas que capturan la información que produce el error gramatical en un nivel alto del procesamiento), pero para ello es requisito imprescindible realizar una muy costosa computación en términos de rasgos y estructuras lingüísticas consolidadas.

Estos sistemas, normalmente, tratan desviaciones puramente gramaticales y/o de estilo y no intentan verificar los errores ortográficos o los llamados errores *tipográficos* ([Kettunen (1996)])[4]. Si se incluyera alguno de estos últimos procesos, se realizaría como paso previo a la pura verificación gramatical, por lo que parece que implícitamente se establece una línea divisoria entre los procedimientos de alto nivel —los de verificación gramatical— y los de bajo nivel —aquellos que actúan en el nivel de la palabra y del signo.

El término *alto nivel* se aplica no solo por la constatable complejidad de los mecanismos computacionales necesarios para tratar estos errores, sino, y esto es lo más importante, por el nivel de abstracción y de la información lingüística necesaria para describir completamente las condiciones de error (y proporcionar correcciones, si ese fuera el caso). El término *bajo nivel* debería significar, pues, lo contrario, y, si bien es cierto que la corrección ortográfica puede realizarse mediante listas exhaustivas de formas léxicas plenas y un simple algoritmo de manejo de caracteres, al menos en determinadas circunstancias, algunos signos de puntuación necesitan un alto nivel de abstracción —y también un alto nivel de análisis lingüístico. De igual modo, algunos errores del nivel sintáctico —esto es, que se extienden sobre más de una

1.8em [3]Nótese que preprocesamiento debe entenderse aquí respecto de los mecanismos centrales de ALEP.

1.8em [4]Estos errores 'textuales' incluyen las convenciones sobre símbolos matemáticos y monetarios, y, de manera general, los signos de puntuación y de representación textual.

palabra— podrían considerarse candidatos para un tratamiento de bajo nivel, si se demostrara que aparecen en contextos de contigüidad aptos para ser tratados en términos de n-gramas y el sistema en su conjunto se viera beneficiado de esta estrategia en su robustez y cobertura de error. En este caso, pues, la línea divisoria entre alto nivel y bajo nivel empezaría a ser más borrosa.

2.1 GramCheck

GramCheck presenta un conjunto de procedimientos de alto nivel para la verificación gramatical. Estos procedimientos reflejan la idea de que (a) los errores violan alguno de los niveles de la descripción lingüística y, por tanto, se producen a diferentes niveles del procesamiento, (b) la verificación gramatical ha de realizarse en el mismo nivel en el que se produce la violación, y (c) las técnicas para la detección y diagnóstico de errores pueden diferir dependiendo del tipo de error. Bajo estas ideas se hallaba también la suposición de que el nivel de tratamiento y la estrategia para determinar la condición de un determinado tipo de error[5] podía realizarse de manera independiente de la lengua.

Los errores característicos detectados y corregidos por este demostrador son, pues, errores de concordancia en género y número, tanto intra- como inter-sintagmática (la corrección es guiada por una heurística basada en un conjunto complejo de ponderaciones), y la adición, sustitución y omisión de preposiciones fuertemente regidas en complementos oblicuos y complementos directos (la preposición *a*). GramCheck también efectuaba algunas verificaciones en el nivel de la palabra, especialmente en el tratamiento de los 'errores cognitivos', esto es, errores en la flexión, que implican que un proceso morfológico regular se aplica a un lema que pertenece a un paradigma flexivo irregular. Si se extiende esta idea, podrían capturarse algunas variantes ortográficas, sean o no flexivas, mediante la inclusión de estas formas incorrectas en el lexicón[6]. Por último, los errores relativos a la puntuación, ya fueran lingüísticamente relevantes o puramente 'cosméticos', no se trataron en absoluto.

Mientras que es probable que e esté de acuerdo en que, para capturar un problema en la selección de la preposición regida, es necesaria una buena cantidad de información lingüística, para los errores en el nivel de la palabra, el trabajo bien podría realizarse en una fase previa, lingüísticamente enriquecida, del procesamiento, esto es, en el nivel del preprocesamiento, sin que hubiera necesidad de efectuar ninguna operación computacional sobre las relaciones de dependencia en las que la palabra incorrecta aparece. De esta forma, el analizador se vería liberado de entradas léxicas espurias cuyo objetivo es proporcionar una 'normalización léxica' (morfológica e, incluso, ortográfica).

Para otros tipos de errores, bajar el nivel en el que son detectados podría paracer menos justificable, pero, al menos a primera vista, algunos de ellos se producen en contextos de con-

1.8em [5]Siempre y cuando pueda demostrarse que las condiciones de error pueden considerarse multilingües, lo que seguramente no es cierto.

1.8em [6]Obsérvese que para la mayor parte de estos errores, muy comunes incluso para los escritores/hablantes nativos, sólo hay una remisión posible —la forma correctamente flexionada. Sin embargo, ningún corrector que trabaje con coeficientes de similitud entre formas propondría esta última, que es, potencialmente, muy diferente a la incorrecta. Así, la forma incorrectamente regularizadala de la primera persona del singular del pretérito indefinido del indicativo **andé*, puede sustituirse certeramente por la forma fuerte correcta de pasado *anduve*, donde un corrector ortográfico propondría otras formas (*ande*, *anda*, *ando*, *nadé*, *nade*), todas ellas más 'cercanas' a la forma incorrecta, pero fracasaría en la captura de la falta de conocimiento que se esconde detrás de esta secuencia intencionada —y tenemos fuertes razones para creerlo– de caracteres.

tigüidad tan estrechos que muy bien podrían formularse en términos de n-gramas. GramCheck, construido sobre una gramática de unificación altamente especificada, que incluye extensiones al mecanismo tradicional de unificación de estructuras de rasgos (los llamados *Constraint Solvers, CSs*), que permiten realizar la relajación de rasgos y las operaciones heurísticas para el diagnóstico y la corrección, podría concentrarse en errores realmente de alto nivel —sintagmáticos, pero no necesariamente violaciones gramaticales contiguas. Aquellas otras situaciones de error en las que pudiera capturarse el elemento incorrecto examinando simplemente su contexto más inmediato podrían ser candidatas para un tratamiento durante la fase de prepocesamiento del analizador.

Existe también otro motivo más para intentar bajar el nivel de detección de ciertos errores: el de la reutilización de las técnicas implementadas en GramCheck, que es sólo posible después de realizar ciertas modificaciones (pocas, pero complejas) en la declaración de ciertos rasgos y en la gramática que se vaya a utilizar (para más detalles técnicos véase [Ramírez y Sánchez-León (1996a)]), lo cual puede entrar en conflicto con las metodologías de implementación. Dada la diversidad de aproximaciones a la implementación de gramáticas, puede parecer que el mejor método para proporcionar estrategias de tratamiento de errores, independientes de la implementación de la gramática, es realizar el mayor número posible de verificaciones de secuencias claramente erróneas durante el prepocesamiento manejando, por un lado, la escueta, pero enriquecible, información disponible en esos niveles, y, por otro, operadores sencillos de búsqueda de patrones de error[7].

Para detectar las condiciones de error, las palabras deben proveerse, al menos, de su información morfosintáctica. Una decisión preliminar que ha de tomarse es si una secuencia de descripciones morfosintácticas (posiblemente ambiguas) pueden proporcionar puntos de anclaje para la detección gramatical. La respuesta es sí en algunos casos y no en otros.

El español es una lengua en la que la primera y/o tercera persona del singular del presente de indicativo de las formas verbales presenta en ocasiones homografía con sustantivos en singular[8]. Además, los artículos son homógrafos de los pronombres proclíticos, de tal forma que, por ejemplo, la secuencia *la tira* puede analizarse igualmente como `pronombre + verbo` y `artículo + sustantivo`. Cualquier cambio en el número del primer elemento daría como resultado una secuencia analizable solamente como `pronombre + verbo`, como es el caso de *las tira*, bajo la suposición de que el texto no contiene errores gramaticales (y la concordancia es uno de ellos).

Sin embargo, nosotros asumimos precisamente que el texto contiene errores y que debe realizarse algún proceso de desambiguación antes de cualquier detección de error. Existe una gran cantidad de trabajos realizados en etiquetado libre de textos con etiquetadores tanto

1.8em [7]Las ideas y técnicas que aquí se proponen pueden utilizarse no solo durante la fase de preprocesamiento de textos en analizadores sintácticos de alto nivel de abstracción (sea en ALEP o en cualquier otra arquitectura), sino también para la confección de correctores gramaticales de bajo nivel que no utilicen verdadero conocimiento lingüístico (profundo). Parte de las ideas aquí vertidas son la base del proyecto *Con-Text*, subvencionado por la Consejería de Educación y Cultura de la Comunidad de Madrid, ref. 05C/002/96, cuyo objetivo es sentar las bases de un verificador morfosintáctico del español en entorno MSDOS-Windows.

1.8em [8]En un lexicón de formas plenas derivado a partir de alrededor de 40.000 entradas, esto ocurre en más de 3.500 sustantivos. Este es solamente uno de los casos 'productivos' de homografía. Hay también muchos homógrafos entre Adjetivos-Verbos y Adjetivos-Sustantivos. Lo que importa es que el español muestra un relativo orden libre en comparación con el inglés. Por ejemplo, muchos adjetivos pueden aparecer en posición pre- y post-nominal y los sintagmas nominales en posición argumental pueden o no seguir al verbo.

estocásticos como basados en reglas. Una precisión aceptable (quizá una marca imbatible) para estos sistemas con textos en inglés se establece entre 95–96% para los etiquetadores estocásticos y 99.7% para los basados en reglas[9]. El trabajo realizado sobre el español en el proyecto CRATER ([Sánchez-León y Nieto-Serrano, (*en prensa*)]) muestra una precisión similar para etiquetadores HMM pero, también aquí, se asume que el texto no contiene errores.

Por último, si aceptamos los resultados tal como se obtienen de un etiquetador[10], el porcentaje de error es una fuente potencial de degradación de la precisión en la detección de errores gramaticales. Ya que esto es algo que nosotros (y seguramente el usuario) no aceptaríamos, podríamos proponer una aproximación más conservadora, según la cual en el nivel del preprocesamiento se intentaría detectar solamente aquellos errores cuya condición es inambigua y su contexto puede establecerse en términos de patrones (semi)fijos.

3 Una aproximación a la verificación gramatical de bajo nivel

Para tener una idea clara sobre la materia que un sistema de corrección gramatical debe abordar, debería realizarse un análisis exhaustivo de textos reales con el fin de extraer de ellos los errores tanto de competencia como, y esto es lo más importante, de actuación que cometen los escritores nativos y los contextos en los que estos aparecen. El resultado de este análisis podría también ayudar a encontrar el equilibrio adecuado para los usuarios entre precisión y cobertura de los sistemas de corrección.

No obstante, con objeto de ilustrar la discusión, presentamos un pequeño conjunto de oraciones y sintagmas extraídos de datos reales[11]:

1. El ejército japonés de Manchuria se desplazó hasta **este**$_{dem\_masc}$ área$_{noun_fem}$.
2. En la$_{det_sing}$ tierras$_{noun_pl}$ bajas$_{adj_pl}$ junto a la costa.
3. Sin embargo, los alemanes refrenaron el avance de las tropas aliadas hasta que **los**$_{det\_masc}$ unidades$_{noun_fem}$ alemanas$_{adj_fem}$ se retiraron.
4. 1.599 liras$_{noun_pl}$ **equivalía**$_{verb\_sing}$ a 1 dólar.
5. La$_{det_sing}$ gente$_{det_sing}$ **vieron**$_{verb\_pl}$ en nosotros nuevas cosas.
6. Son$_{verb_pl}$ **falsa**$_{adj\_sing}$ y hay que probarlas.
7. Al menos 37 ciudadanos negros de Sudáfrica resultaron$_{verb_pl}$ **muerto**$_{adj\_sing}$ el pasado fin de semana en los enfrentamientos étnicos.

1.8em [9]La bibliografía sobre Modelos markovianos (ocultos) (*Hidden Markov Models*) siempre muestra resultados referidos a corpus específicos, usando un etiquetario particular (más o menos fino) y un lexicón exhaustivo o podado de ambigüedades poco productivas, lo que dificulta hacer una comparación entre los distintos sistemas de etiquetado. Con respecto a aproximaciones basadas en reglas, [Karlsson *et al.* (1995)] proporcionan los resultados antes mencionados utilizando *Constraint Grammars* (CGs), un formalismo de desambiguación morfosintáctica basado en FSAs, pero aún permiten un 3–7% marginal de palabras ambiguas.

1.8em [10]Hay que hacer notar que los análisis gramaticales que integran los resultados de un etiquetador necesitan otro tipo de verificación, la de los errores producidos por el propio etiquetador (véase [Elworthy (1994)]).

1.8em[11]Estos datos han sido extraídos de una versión beta de la edición española de la enciclopedia *Encarta97* de Microsoft, de *El periódico mensual humanista del barrio de Lista* (Abril 1996, no. 14), y *El País* (8/4/96 y 9/12/96).

8. Se estableció una jerarquización de los daimios **de acuerdo a** sus relaciones con los Tukugawa.

9. Hay empresas que se comprometen **ha**$_{verb\_pres}$ hacer$_{verb_inf}$ la reforma.

10. La ingente descentralización posterior se ejecutó **por Suárez y Felipe González** sin contar con AP.

11. Vieron **como** el sector perdía importancia.

Estas oraciones presentan diferentes tipos de error, pero todas ellas comparten el hecho de que el elemento incorrecto podría capturarse verificando su contexto más inmediato. Este contexto es lo que podríamos denominar su condición de error. La condición de error ha de entenderse como una secuencia de formas léxicas inambiguas (morfosintácticamente), que las reglas de captura de error del nivel del preprocesamiento manejarían como si fueran patrones fijos. Utilizando información morfosintáctica y algún conocimiento léxico, deberíamos poder tratar errores de concordancia, inadecuaciones léxicas y quizá otras secuencias incorrectas[12]. Las reglas de *patrones morfosintácticos*, por un lado, detectarían secuencias erróneas de palabras contiguas y las reglas de *patrones léxicos*, por otro, detectarían inadecuaciones en el nivel de la palabra, tales como errores cognitivos.

La posibilidad de realizar verificaciones fiables que impliquen una cierta abstracción depende en gran medida de las funcionalidades del etiquetador que utilicemos. El etiquetador mpro utilizado en el módulo alemán del proyecto LSGRAM, proporciona información morfosintáctica de cada uno de los elementos del texto que etiqueta, pero también es capaz de construir, de forma superficial, las relaciones sintagmáticas que se establecen entre las formas léxicas (véase para más información [Declerck y Maas (1997)]). Esta información es más que suficiente para formular reglas de patrones morfosintácticos de estructura compleja, que permitan la veficación, por ejemplo, de la concordancia en elementos no necesariamente contiguos. Sin embargo, esta no suele ser la salida de la mayoría de los etiquetadores, cuya labor es segmentar y proporcionar, tras la desambiguación, una etiqueta morfosintáctica para cada una de los elementos de un texto. En este contexto, nuestras reglas morfosintácticas solo pueden ser formulaciones de secuencias de palabras (*n-gramas*) que, en el fondo, mostrarían (parte de) las relaciones sintagmáticas de las formas léxicas contiguas, si es que las hubiera. Sin embargo, formuladas de este modo, aquellas relaciones sintagmáticas que no aparecieran en contextos contiguos podrían provocar la sobredetección de errores. Deberíamos, pues, encontrar algún mecanismo para poder verificar esos elementos no contiguos. Para ello, sería necesaria información estructural más profunda, como, por ejemplo, información sobre coordinación o funciones sintácticas[13].

1.8em

[12]Nótese que algunos de los errores que aparecen en nuestras oraciones no podrían tratarse en los términos que proponemos, ya que en la situación de error aparece una forma léxica que presenta homografía con otras. Este es el caso de *como* en la oración 11 ('como$_{verb_1pers_sing}$', 'como$_{adv}$', 'como$_{causal_conj}$', 'como$_{conditional_conj}$').

1.8em

[13]Piénsese, por ejemplo, en casos como *dos*$_{card\_pl}$ *mesas*$_{noun\_pl}$ *de*$_{prep\_of}$ *madera*$_{noun\_sing}$ *verdes*$_{adj\_pl}$ y *una*$_{det\_sing}$ *mesa*$_{noun\_sing}$ *y*$_{conj}$ *una*$_{det\_sing}$ *silla*$_{noun\_sing}$ *verdes*$_{adj\_pl}$. Si una regla de captura de error estableciera como condición de error que un sustantivo singular no debe ir seguido de un adjetivo plural, *madera verdes* y *silla verdes* serían detectadas como secuencias incorrectas, lo que no es deseable en estos contextos. Así pues, la información estructural es necesaria para describir condiciones inambiguas de error. Dependiendo de las funcionalidades del etiquetador que utilicemos, las posibilidades de detección de errores a bajo nivel pueden crecer en fiablidad. Para obtener

Un aspecto importante de la verificación gramatical en el nivel del preprocesamiento es que esta se limitaría simplemente a la detección de errores. En ningún momento se esperaría en este nivel la corrección del error o una simple sugerencia sobre cómo corregirlo. En este sentido, las reglas de captura de error del nivel del preprocesamiento se ocuparían de la detección, por ejemplo, de la falta de concordancia en elementos contiguos, contrastándose únicamente los valores de género y número de tales elementos y dejándose a un lado comprobaciones más complejas como la averiguación de si el sustantivo implicado tiene género inherente o no, lo que sería importante si nuestro objetivo fuera guiar la corrección del error, ya que necesitaríamos conocer cuál es el elemento o elementos erróneos.

Si bien la detección de errores de concordancia presenta ciertas dificultades si la salida de nuestro etiquetador tiene el formato habitual, es posible encontrar otros contextos fiables durante esta fase temprana del procesamiento en algunos patrones prácticamente fijos, formados por palabras que, bien porque pertenecen a un paradigma flexivo irregular, se les aplica una regularización errónea, bien porque muestran una irregularidad superficial en determinados contextos, esta se aplica erróneamente a otros contextos, produciéndose, entonces, el error. Este es el caso del patrón formado por un artículo masculino seguido de un sustantivo de género femenino débil que comienza por a acentuada (los hablantes nativos tienden a asimilar el comportamiento del resto de los determinantes al del artículo (oración 1)). La condición de error que presenta esta secuencia es fiable, ya que solamente en el caso de que dicho sustantivo sea singular y esté precedido de un artículo, este ha de ser masculino. En cualquier otro caso, los premodificadores (incluidos el artículo plural, demostrativos, cuantificadores, adjetivos, etc.) y postmodificadores han de ser femeninos.

Esta situación de error es bastante simple y puede ser formulada mediante una regla de patrón erróneo de bajo nivel basado en *n-gramas* de descripciones morfosintácticas, aunque el error sea de alto nivel (ya que presupone una cierta abstracción del conocimiento lingüístico). Otros errores pueden detecterse del mismo modo. Las amalgamas son un ejemplo de esta clase. El error se produce, en este caso, cuando la amalgama es descompuesta en las preposiciones *a* o *de* y el artículo determinado masculino *el* (**a el*, **de el*). Del mismo modo, errores inintencionados que producen las secuencias de una amalgama junto a otro artículo definido (**del el*, **del la*) pueden detectarse mediante patrones fijos. Debe hacerse notar que este tipo de error no es detectado por los correctores ortográficos, ya que ambas formas son correctas. A este respecto, también es posible realizar una normalización textual si se detectara la falta de artículo en determinados nombres propios en un contexto de amalgama, como el nombre propio *El Cairo* en secuencias como **la población del Cairo*[14] (i.e. *la población de El Cairo*).

Otras secuencias erróneas, como elementos interpuestos entre el auxiliar y el participio en los tiempos compuestos verbales, la forma *se* (por *sé*) seguida de un signo de puntuación, de una conjunción o sustantivo, verbos copulativos en plural seguidos de adjetivos en singular, la

---

esa fiabilidad, otro tipo de información como la textual (e.g. principio de oración) puede ser también útil para delimitar el contexto de aplicación de una regla. Una regla que utilizara información textual podría capturar los errores de las oraciones 4 y 5, donde un SN (i.e. el sujeto) aparece al comienzo de la oración precediendo al verbo, y donde no hay concordancia entre dicho SN y el verbo. Las condiciones textuales, pues, describen que no hay ningún elemento coordinado que pudiera hacer pensar que el verbo ha de estar en plural. No obstante, incluso delimitando el contexto hasta este punto, sería necesaria información semántico-léxica que determinara qué tipo de sustantivos pueden ser sujeto de estos verbos, rechazándose, por ejemplo, los sustantivos locativos y temporales. Aun así, la fiabilidad de nuestra regla no sería total. Por el contrario, el error de la oración 6, donde un verbo copulativo no concuerda con el adjetivo que le sigue podría detectarse de manera directa.

[14] Habría, sin duda, que hacer una distinción entre nombres propios que obligatoriamente aparecen con artículo y los que presentan opcionalidad.

conjunción disyuntiva *o* ante una palabra que comience por o, etc., son algunos ejemplos de secuencias erróneas que podrían ser detectadas mediante reglas de patrones morfosintácticos o grafémicos. Del mismo modo, ciertos errores en signos puntuación, como la falta de uno de los signos de puntuación balanceados —interrogaciones, paréntesis, etc.—, falta de punto final en algunas abreviaturas, espacios antes y después de los paréntesis, los guiones, o falta de espacio después de un signo de puntuación, son también ejemplos de errores de caracter tipográfico que podrían ser detectados en el nivel del preprocesamiento, liberando de este modo las fases posteriores de un trabajo costoso que es fácilmente tratable, y poco costoso, a este nivel.

La normalización textual que podría llevarse a cabo, pues, durante la fase de preprocesamiento también está relacionada con la normalización 'léxica'. Durante esta fase, pueden detectarse errores cognitivos (morfológicos u ortográficos) por medio de la inclusión de las formas incorrectas en el lexicón de este nivel. Inadecuaciones léxicas, errores (casi) sistemáticos debidos a una falta de conocimiento, errores ortográficos e incluso referencias (e.g. *de acuerdo con* es preferible a *de acuerdo a* (oración 8 *supra*))[15] pueden ser capturados al nivel de la palabra, que es realmente el nivel al que pertenecen tales inadecuaciones.

La situación más común respecto de los errores cognitivos implica el cambio por parte de los hablantes/escritores nativos del paradigma flexivo irregular al que pertenece un lema a un paradigma flexivo regular. En este caso, las regularidades incorrectas pueden asociarse de manera casi automática a las irregularidades correctas:

- plural:
  *menú–menús/*menúes* ,
  *campus–campus/*cámpuses*,
  *convoy–convoyes/*convoys*;

- pasado fuerte vs. pasado débil:
  *traducir–tradujiste/*traduciste*,
  *andar–anduve/*andé*,
  *caber–cupo/*cabió*;

- errores morfológicos cognitivos de tipo ortográfico:
  *espurio–*espúreo*.

Los patrones de error que hemos analizado aquí podrían parecer ocasionales, pero las posibilidades reales de la verificación a bajo nivel dependen de los resultados de un trabajo empírico que manejara grandes cantidades de texto, lo que nos revelaría los errores reales que pueden ser tratados en este nivel y los que deben ser tratados en niveles superiores.

4 Conclusiones

Este artículo presenta algunas reflexiones sobre la verificación gramatical durante el nivel del preprocesamiento. Se han propuesto, de manera aproximativa, técnicas y procedimientos para

1.8em [15]Trabajar con referencias no es una tarea sencilla y en ocasiones se necesitaría información semántica, ya que la referencia puede dirigirse hacia una de las acepciones de la entrada léxica. Este es el caso de *colacionar* para el que el *Diccionario de la Real Academia* especifica la referencia *cotejar* cuando significa 'referir'. Otras referencias son más simples de tratar, como *balompié*, cuya forma preferida es *fútbol*.

tratar textos incorrectos a este nivel, previo al procesamiento de alto nivel, con la finalidad de que sirvan como punto de partida para una discusión sobre robustez. Con este objetivo, es necesario también prever mecanismos simples e independientes de la implementación de las gramáticas. Pero también hemos visto que, contrariamente a lo que se pensaba en un principio, la verificación a bajo nivel no tiene las mismas posibilidades en todas las lenguas, ya que hay que tener en cuenta factores como la ambigüedad léxica y contextual. No obstante, es urgente la realización de un estudio comparativo en diferentes lenguas con objeto de definir una estrategia común sobre el tratamiento de errores en textos etiquetados.

Referencias

[Adriaens (1994)] Adriaens G. 1994. The LRE SECC Project: Simplified English Grammar and Style Correction in an MT Framework, In *Proceedings of Language Engineering Convention.* pp. 1–8.

[Bolioli *et al.* (1992)] Bolioli A., L. Dini, G. Malnati. 1992. JDII: Parsing Italian with a Robust Constraint Grammar, In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92).* pp. 1003–1007.

[Bredenkamp *et al.* (1996)] Bredenkamp, A., T. Declerck, F. Fouvry, B. Music. 1996. Efficient Integrated Tagging of Word Constructs, In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-96).* pp. 1028–1031.

[Declerck y Maas (1997)] Declerck, T., D. H. Maas. 1997. The Integration of a Part-of-Speech Tagger into the ALEP Platform, In *Proceedings of the 3rd AUG Workshop.*

[Elworthy (1994)] Elworthy, D. 1994. Automatic Error Detection in Part of Speech Tagging, In *Proceedings of the International Conference on New Methods in Language Processing.* pp. 190–195.

[Heidorn *et al.* (1982)] Heidorn G. E., K. Jensen, L. A. Miller, R. J. Byrd, M. S. Chodorow. 1982. Devoloping a Natural Language Interface to Complex Data, In *ACM Trans. on Database Sys.*, 3(2). pp. 105–147.

[Genthial *et al.* (1994)] Genthial D., J. Courtin, J. Ménézo. 1994. Towards a more user-friendly correction, In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94).*

[Karlsson *et al.* (1995)] Karlsson, F., A. Voutilainen, J. Heikkilä, A. Anttila (*eds.*). 1995. *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter, Berlin - New York.

[Kettunen (1996)] Kettunen, K. 1996. Low-level Typographical Spellchecking: A proposal, In *Computers and the Humanities*, 30(1). pp. 77–84.

[Maas (1996)] Maas, H. D. 1996. MPRO – Ein System zur Analyse und Synthese deutscher Wörter, In Hausser Roland, editor, *Linguistische Verifikation, Sprache und Information Nr. 34*, Max Niemeyer Verlag, Tübingen.

[Maas y Hirschfeld (1996)]. Maas, H. D., D. Hirschfeld. 1996. Improving the Functionality of a Text-to-Speech System by Adding Morphological Knowledge, In Görz Guenther y Hoelldoblerund Steffen, (*eds.*), *KI-96: Advances in Artificial Intelligence*, Springer (Lecture Notes in Artificial Intelligence 1137).

[Ramírez y Sánchez-León (1996a)] Ramírez Bustamante, F., F. Sánchez-León. 1996. Is linguistic Information enough for grammar checking?, In *Proceedings of the First International Workshop on Controlled Language Applications, CLAW '96.* pp. 216–228.

[Ramírez y Sánchez-León (1996b)] Ramírez Bustamante, F., F. Sánchez-León. 1996. GramCheck: A Grammar and Style Checker, In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96).* pp. 175–181.

[Sánchez-León y Nieto-Serrano, (*en prensa*)] Sánchez-León, F., A. Nieto-Serrano. *forthcoming*. Retargeting Taggers, In R. Garside, G. Leech y A. M. McEnery (*eds.*), *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Longman, London.

[Schmidt et al. (1996)] Schmidt, P., S. Rieder, A. Theofilidis, T. Declerck. 1996. Lean formalisms, Linguistic Theory and Applications. Grammar Development in ALEP, In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96).* pp. 286–291.

[Thurmair (1990)] Thurmair G. 1990. Parsing for grammar and style checking, In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-90).* pp. 365–370.

[TWB (1992)] Translator's Workbench. 1992. Final Report of the Esprit Project 2315.

[Vosse (1992)] Vosse, T. 1992. Detecting and Correcting Morpho-syntactic Errors in Real Texts, In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ACL-92).* pp. 111–118.