# GramCheck: Un corrector gramatical para español

Flora Ramírez Bustamante
Fernando Sánchez León
Laboratorio de Lingüística Informática
Facultad de Filosofía y Letras
Universidad Autónoma de Madrid
flora,fernando@maria.lllf.uam.es

**Abstract**

This paper presents a grammar and style checker demonstrator for Spanish native writers developed within the project GramCheck. Besides a grammar and style error typology for Spanish, a linguistically motivated approach to detection and diagnosis is presented. The demonstrator includes full coverage for agreement errors and certain head-argument relation issues. It also provides correction by means of an analysis-transfer-synthesis cycle.

## 1 Introducción

GramCheck es un demostrador de un sistema de corrección gramatical para hablante nativos de español y griego moderno. Este sistema se ha desarrollado en el marc del proyecto del mismo nombre, MLAP93-11, que ha sido cofinanciado por la Uniói Europea en la convocatoria *Multilingual Action Plan*. El proyecto, que ha durad dieciocho meses, ha sido realizado por un consorcio coordinado por la Universida Carlos III de Madrid, en el que también figuran la empresa griega KNOWLEDGE S.A. y el grupo editorial español ANAYA (en concreto, la editorial Biblograf, S.A. como usuario piloto.

El demostrador ha sido implementado sobre la plataforma para desarrollo de apli caciones de ingeniería lingüística creada por la Unión Europea, ALEP [ET6/1, 1991] [Simpkins, 1994]. Utiliza la arquitectura cliente-servidor propio del sistema X Win dows, Motif como gestor de ventanas y Xminfo como formato de almacenamiento de l base de conocimiento. GramCheck hace un uso generalizado de ciertas extensiones a formalismo de rasgos con tipos, basado en unificación, implementado en ALEP. Esta extensiones (llamadas *Constraint Solvers*, CSs) son fragmentos de código en PROLO que permiten realizar diversas operaciones booleanas y relacionales sobre valores d rasgos.

Asimismo, GramCheck se ha beneficiado de los resultados parciales de LS-GRAM (LRE-61029), otro proyecto europeo cuyo objetivo es la implementación de gramáticas de cobertura intermedia para varias lenguas europeas [Badia *et al.*, 1995].

El demostrador, en su versión actual, comprueba si un documento contiene errores gramaticales o ciertas deficiencias de estilo, y, en caso afirmativo, proporciona al usuario mensajes, sugerencias y una corrección automática. Además, el demostrador incluye un primer tratamiento para fenómenos específicos de ciertos sublenguajes, por lo que es posible elegir entre variedades de lengua estándar y administrativa[1].

Este artículo presenta una descripción de la tareas más importantes llevadas a cabo en el proyecto GramCheck en relación con la identificación y clasificación de los errores más comunes cometidos por hablantes nativos en textos escritos y la integración de técnicas que permitan detectar, diagnosticar y, en la medida de lo posible, corregir dichos errores sin renunciar al uso de mecanismos de PLN ampliamente aceptados. Finalmente, se incluyen algunos datos sobre errores tratados (o tratables) con las técnicas propuestas y errores no tratados (y no tratables) con dichas técnicas.

## 2 Tipología de errores

Para la elaboración de la cobertura del corrector, se examinaron documentos editados (procedentes de artículos de periódicos, libros de diversa índole), orales[2], primeras versiones y versiones revisadas de textos escritos por diferentes profesionales (de la informática y sociología), así como cartas comerciales. El corpus final contenía alrededor de 70.000 palabras además de 90 cartas comerciales.

GramCheck distingue, en principio, entre errores gramaticales y deficiencias de estilo, siendo la única diferencia funcional entre ambos el hecho de que para los primeros se proporciona, junto a las explicaciones y sugerencias relacionadas con el problema lingüístico detectado, una corrección.

Sin embargo, el examen de este corpus reveló que, desde el punto de vista del análisis, era mucho más interesante una distinción entre errores que violan restricciones estructurales y errores que violan restricciones no estructurales. Así, utilizando ambos parámetros, podemos establecer la siguiente tipología de errores[3]:

1. Errores gramaticales:

   (a) Los errores no estructurales afectan a:

   - la concordancia intra-sintagmática en género y número, como entre un sustantivo y sus modificadores en posición prenominal y postnominal

[1] A este fin, el módulo de análisis se ha concebido como una gramática nuclear y dos subgramáticas satélites —una para el sublenguaje administrativo y la otra para los casos que resultan en un solapamiento entre éste y el lenguaje estándar— que son mutuamente excluyentes. Consecuentemente, la activación de una subgramática implica la desactivación de la otra, y, en ambos casos, dependiendo de la selección del usuario, la subgramática se incorpora a la gramática nuclear.

[2] Ambos proporcionados por ANAYA.

[3] Una tipología de errores más refinada puede verse en [Ramírez *et al.*, 1994].

(determinantes, cuantificadores, adjetivos, etc.): **Algunos servicios van a ver modificados sus horario.*

- la concordancia inter-sintagmática en persona, número, género y caso entre dos o más constituyentes dentro de la oración, como concordancia entre sujeto-verbo (activo o pasivo) o sujeto/objeto-atributo: **Este estilo caracterizan el centro.*

(b) Los errores estructurales afectan a:

- las relaciones entre núcleo y argumento, como la omisión, adición o sustitución de una preposición subcategorizada: *Se acordó *(de) que tenía una reunión por la mañana,*
- ciertos modificadores, como la sustitución de una oración subordinada relativa por un gerundio: **La carta conteniendo los datos,*
- la adición de sujeto en las impersonales con *se*: *Los ordenadores se usan (*por los alumnos),*
- la omisión, adición y sustitución de caracteres: *La empresa que ha realizado el desarrollo desea *informa del proyecto, En principio, podría *parecen una DB orientada a objetos.* A esta clase pertenecen la adición o (sobre todo) omisión de acentos, lo que provoca confusión entre pronombres interrogativos y conjunciones subordinadas, por ejemplo: *Se ha visto *como las consultas al sistema se transfieren de un nivel otro*, la adición, sustitución u omisión de signos de puntuación, como comas después de ciertos transconstruccionales, el cierre de signos balanceados (paréntesis, comillas, corchetes, etc.) o la adición de coma no justificada, entre sujeto y verbo, o verbo y argumentos en posición postverbal, y la adición u omisión de espacios en blanco entre palabras: *Los dólares significan algo con respecto *así mismos*[4].

2. Deficiencias de estilo:

(a) En el corpus aparecen deficiencias estructurales relacionadas con calcos sintácticos de otras lenguas, como la construcción sustantivo+*a*+infinitivo: *Los productos a entregar se relacionan a continuación.*

(b) Las deficiencias léxicas (no estructurales) son de muy variada índole:

- Uso de palabras extranjeras en lugar de sus equivalentes españolas: *handicap* en lugar de *desventaja.*
- Falsas derivaciones: *concretizar* en lugar de *concretar.*
- Expresiones latinas, como *de motu propio* en lugar de *motu proprio.*

[4]En este tipo se engloban solamente los errores que resultan en una secuencia de caracteres (palabra) existente en la lengua y, por tanto, que escapa al dominio de verificación de los correctores ortográficos comerciales. Buena parte de estos errores se encuentran, con toda seguridad, en el ámbito de la corrección tipográfica (u ortográfica), si bien, por el motivo anterior, es necesario prever un tratamiento para ellos en un módulo gramatical.

- Unidades multipalabra que pueden sustituirse por otras simples, como *en el momento actual* en lugar de *actualmente*.

(c) Abuso de ciertas construcciones, como la pasiva canónica o las formas de gerundio, etc.

El número total de errores identificado en el corpus es de 543. Dada la ausencia de otras fuentes de información, la estadística que presentamos en la tabla 1 ha de verse como una frecuencia representativa de errores en textos escritos por hablantes nativos. En esta tabla, se presentan separados los errores que tienen una motivación lingüística (o que son susceptibles de una interpretación lingüística del error) de aquellos que, siendo —seguro— más fortuitos que los primeros, son solamente "descuidos" tipográficos. Así, los errores de concordancia y los estructurales representan genuinamente el ámbito de la corrección gramatical, mientras que las deficiencias de estilo representan el ámbito de la detección de problemas de estilo.

Tabla 1: Estadística de errores

| Tipo de error | | Porcentaje |
|---|---|---|
| Errores de concordancia | | 18.5 |
| Errores estructurales | | 9.7 |
| Puntuación | Omisión | 32.2 |
| | Adición | 4.8 |
| Errores en el nivel léxico | Caracteres | 6.3 |
| | Acento | 8.0 |
| Deficiencias de estilo | Estructurales | 3.5 |
| | Léxicas | 12.0 |
| Otros | | 5.0 |

Estos datos, especialmente los que conciernen a los errores de concordancia, contradicen las afirmaciones de algunos desarrolladores de correctores gramaticales. En este sentido, [Veronis, 1988] afirma que los escritores nativos no cometen errores relacionados con rasgos morfológicos. Por el contrario, [Vosse, 1992] los acepta, a pesar de que un examen de textos realizado por este autor parece revelar que no son frecuentes en textos escritos por nativos. Ambos coinciden en caracterizar los errores morfosintácticos como un ejemplo de falta de competencia. Sin embargo, nuestro corpus muestra que casi una quinta parte de los errores son de este tipo.

# 3 Técnicas

La estrategia global para la detección, el diagnóstico y la corrección de errores gramaticales y deficiencias de estilo se vertebra sobre los siguientes parámetros:

- Para la detección, se ha adoptado una combinación de la técnica de relajación de rasgos y la de anticipación de errores. La primera técnica se ha implementado utilizando CSs, mientras que para la segunda se han implementado reglas explícitas adecuadamente definidas en la gramática nuclear o en las subgramáticas satélite. Los CSs permiten la relajación de ciertos rasgos en las reglas de la gramática cuya unificación se decide posteriormente cuando se evalúa el CS. De este modo las reglas no realizan la comprobación de valores, por lo que los CSs desempeñan un papel crucial al realizar una unificación de variables extendida que permite adoptar decisiones consecuentemente.

  No obstante, la novedad de GramCheck es que ciertos errores, susceptibles de un tratamiento mediante anticipación, se han reanalizado como violaciones de rasgos y, por tanto, pueden tratarse por medio de la relajación de ciertos rasgos que codifican, doblemente para estos casos, tanto la información correcta como la incorrecta (aunque probable) en la entrada léxica. Este reanálisis afecta a los valores de algunos rasgos, como los de la preposición regida, que ha sido implementada como una descripción de patrones que asocian el patrón incorrecto (o los patrones incorrectos) al patrón correcto.

  Esta técnica de relajación de patrones relacionados está lingüísticamente motivada en la constatación de que los escritores nativos sustituyen, por ejemplo, una preposición por otra cuando existe una asociación entre patrones que muestran las mismas propiedades léxico-semánticas y/o sintácticas. Piénsese, así, en el uso del adjetivo comparativo *inferior* junto a *que*, en lugar de la preposición *a*. Este error es debido a la relación inconsciente que se establece entre el patrón de la comparación *más... que* y la forma comparativa de esta unidad léxica. Computacionalmente, la técnica de patrones relacionados ha de verse como un medio de control de la relajación de las restricciones estructurales que impide la explosión de múltiples análisis.

- El diagnóstico de los errores se realiza mediante CSs y una técnica heurística (similar a la propuesta por otros autores [Veronis, 1988], [Bolioli *et al.*, 1992] [Vosse, 1992], [Genthial *et al.*, 1994]), para los errores de concordancia, y la ya mencionada técnica de patrones relacionados, para algunos de los errores estructurales (como los de preposición regida).

  Para la comprobación de la concordancia, todos los elementos que intervienen en ella reciben un peso heurístico inicial. Al comprobar los valores de los rasgos involucrados en la concordancia, se suman todos los pesos de los elementos que concuerdan, con lo que se obtiene, cuando se culmina el proceso de análisis, un peso máximo, que se asocia al valor de un rasgo de carácter morfosintáctico.

  La idea en la que se basa esta heurística es que dependiendo de una serie de principios lingüísticos basados en propiedades morfosintácticas, los valores de género y número de algunas unidades léxicas tienen un valor lingüístico más grande que los valores de otras unidades, por lo que hay que asignarles un peso mayor. La heurística, pues, se basa en la parametrización de dos tipos de información:

1. el constituyente que contiene los valores que en una situación de error controlan el resto de los valores de otros constituyentes (como, por ejemplo, el género de los sustantivos con género inherente en oposición al de aquellos sustantivos con moción de género, o el número de los cardinales diferentes de *uno*, que han de controlar el resto del sintagma),
2. la evaluación del número de constituyentes que comparten o no los mismos valores.

- La corrección se basa en la transferencia (*transfer*) de las Estructuras Lingüísticas (EL) que contienen información errónea a una "lengua" definida como español correcto. Estas estructuras se sintetizan luego con el fin de ser mostradas al usuario. El diseño, pues, es similar al de un sistema de traducción automática basado en transfer[5], siendo la mayor diferencia entre ambos el que GramCheck incorpora los mecanismos mencionados para la corrección y el hecho de que no todas, sino solamente aquellas estructuras que contienen error, completan el ciclo de transfer y generación.

# 4 Cobertura de errores y limitaciones del sistema

GramCheck ha demostrado la madurez de ALEP como plataforma para el desarrollo de aplicaciones de PLN, de las que la correción gramatical es solo un ejemplo. Asimismo, se han propuesto una serie de técnicas para incorporar la robustez necesaria para analizar oraciones correctas y mal formadas. Sin embargo, la cobertura de la gramática sobre la que se fundamenta el sistema es insuficiente para pensar en una aplicación industrial del demostrador. En este sentido, su cobertura debería ampliarse con objeto de conseguir un producto operativo y fiable.

Sin embargo, incluso con una gramática y un lexicón mayores, capaces de proporcionar análisis para un alto porcentaje de las oraciones de un texto, es necesario ampliar los tipos de errores tratados, puesto que, en un análisis optimista de los errores para los que existe tratamiento, el sistema, en su versión actual, sería capaz de detectar (y corregir) un 28,2% del total de los errores (correspondiente a los errores de ámbito estrictamente gramatical) más otro 15,5% (correspondiente a los problemas de ámbito estrictamente estilístico). Esto significa que todavía existe un 56,3% de errores para los que no hay tratamiento. Por tanto, un desarrollo posterior de GramCheck en términos de cobertura gramatical (y obviando otras cuestiones, también importantes, como la segmentación de oraciones, la correcta identificación de nombres propios en un texto, etc.) sería capaz de tratar, en el mejor de los casos, sólo la mitad de los errores de un texto[6].

GramCheck, en su versión actual, trata una muestra representativa de los errores de carácter lingüístico en los ámbitos gramatical y de estilo, que incluye:

[5]Como en otros correctores gramaticales para el inglés basados en gramáticas de unificación [Adriaens, 1994].

[6][Ramírez & Sánchez, 1996] incluye un primer esbozo para el tratamiento de otros tipos de errores

- Errores intra- e inter-sintagmáticos (género y/o número en oraciones activas —con verbos predicativos y copulativos— y pasivas).
- Objetos directos: omisión de la preposición *a* junto a entidades animadas y adición de esta preposición en entidades no animadas.
- Adición, omisión y sustitución de preposiciones regidas, incluidos los fenómenos de *dequeísmo* y *queísmo*.
- Errores en amalgamas (*de el* en lugar *del*, por ejemplo).
- Deficiencias estructurales (como el uso de la estructura sustantivo+*a*+infinitivo en el lenguaje estándar, aunque tal construcción es admitida en el sublenguaje administrativo).
- Deficiencias léxicas (de los tipos mencionados anteriormente).
- Uso abusivo de formas pasivas y gerundios.

Con todo, las futuras extensiones de GramCheck se van a centrar, más que en el desarrollo ulterior de la gramática y del lexicón (necesarios en cualquier caso), en el reanálisis de otros fenómenos de error en términos de relajación de rasgos, allí donde sea posible, y en el empleo de otras técnicas de PLN (por ejemplo, el uso de trigramas) para la captura temprana de posibles secuencias de error.

# Referencias

[Adriaens, 1994] Adriaens G. The LRE SECC Project: Simplified English Grammar and Style Correction in an MT Framework, en *Proceedings of Linguistic Engineering Convention*, París, Julio 1994.

[Badia *et al.*, 1995] Badia, T., M. Carulla, M. Melero. Resumen del proyecto LS-GRAM, en *Procesamiento del Lenguaje Natural*, 16:85–88.

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

[ET6/1, 1991] Pulman, S. G. (ed.). *EUROTRA ET6/1: Rule Formalism and Virtual Machine Design Study*, SRI International, Cambridge Computer Science Research Centre, ©Commission of European Communities.

[Genthial *et al.*, 1994] Genthial D., J. Courtin, J. Ménézo. Towards a more user-friendly correction, en *Proceedings of the 16th International Conference on Computational Linguistics (COLING-94)*, Kyoto, Japón.

[Ramírez *et al.*, 1994] Ramírez, F., Y. Rodríguez, F. Sánchez León. Tipología de errores gramaticales del español para un sistema automático de corrección de textos, en *Actas del X Congreso de Lenguajes Naturales y Lenguajes Formales*, PPU, Barcelona 1994.

[Ramírez & Sánchez, 1996] Ramírez, F., F. Sánchez León. Is linguistic information enough for grammar checking?, en *Proceedings of the First International Workshop on Controlled Language Applications, CLAW'96*, Katholieke Universiteit Leuven: 216–224.

[Simpkins, 1994] Simpkins N. K. An Open Architecture for Language Engineering: The Advanced Language Engineering Platform (ALEP), en *Proceedings of Linguistic Engineering Convention*, París, Julio 1994.

[Veronis, 1988] Veronis J. Morphosyntactic correction in natural languages interfaces, en *Proceedings of the 13th International Conference on Computational Linguistics (COLING-88)*: 708–713.

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