UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

TESIS DOCTORAL

FACULTAD DE TRADUCCIÓN E INTERPRETACIÓN

Lingüística de errores con fines computacionales

PatErr, un recurso para la revisión textual del español basado en patrones de error codificados

Autora: Lydia Arroyo Herrero

En Las Palmas de Gran Canaria, en junio de 2017

UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

FACULTAD DE TRADUCCIÓN E INTERPRETACIÓN

Programa de Doctorado: Español y su cultura I+D+i

Lingüística de errores con fines computacionales

PatErr, un recurso para la revisión textual del español basado en patrones de error codificados

Tesis doctoral presentada por Lydia Arroyo Herrero
Dirigida por el Dr. Francisco Javier Carreras Riudavets

El director *La doctoranda*

Las Palmas de Gran Canaria, junio de 2017

Resumen

Esta tesis recoge las aportaciones que ofrece la investigación teórica y la formalización del uso de la lengua en el Procesamiento del Lenguaje Natural. Es, pues, un trabajo multidisciplinar que aúna fuerzas, conocimientos y métodos para la consecución de una empresa común: el desarrollo de un recurso competitivo y viable para la revisión textual automática del español, que se actualiza en un repertorio de patrones de errores codificados con un lenguaje asumible para el entorno computacional.

Para estos programas, según el enfoque en el que se asienta esta propuesta que denominaremos *lingüística de errores con fines computacionales,* el objeto de estudio y tratamiento debe ser el error; de él se extraerán hipótesis, abstracciones, generalizaciones y casos concretos que deberán ser formalizados para que un autómata pueda identificarlos en el texto.

Inserto en esta perspectiva, se ha llevado a cabo el planteamiento, metodología y desarrollo teórico y práctico de este recurso —PatErr—. Tomándolo como núcleo y asociado a herramientas básicas de análisis automático, puede ponerse en marcha un corrector competitivo en el contexto de las industrias de la lengua.

Para estos desarrollos son ineludibles varias tareas previas de adaptación del conocimiento lingüístico al entorno computacional. Dos de las concreciones que han posibilitado este enlace necesario para la consecución de esta propuesta han sido, por un lado, el desarrollo de un etiquetario para las unidades que componen un lexicón de español, y por otro, un lenguaje de codificación expresivo y generativo que ha permitido llevar a cabo con precisión la transferencia de conocimientos.

Abstract

This thesis aims to collect the contributions of theoretical research and formalization of the use of language in an environment of Natural Language Processing. It is, by definition, a multidisciplinary work that combines forces, knowledge and methods of analysis for the achievement of a joint venture: the development of a competitive and viable resource for automatic textual revision of Spanish texts, which is updated in a repertoire of error patterns coded with a language acceptable in a computational environment.

On these programs, according to the approach on which this proposal is based —that we will call *Error Linguistics for Computational Purposes*—, the object of study and treatment must be the error; from it, hypotheses, abstractions, generalizations and concrete cases will be extracted and formalized so that an automaton can identify them in the text.

From this perspective, the approach, methodology and theoretical and practical development of a resource —PatErr— has been performed based on codified error patterns that will have to be identified in the text under review. Taking it as core and associated with basic tools of automatic analysis, a program of textual review competitive in the context of the language industries can be launched.

For these developments, several previous tasks of adaptation of the linguistic knowledge to the computational environment have been necessary. Two of the concretions that have enabled this link have been, on the one hand, the development of a tagger for the units that compose a lexicon of the Spanish language, and on the other, an expressive and generative coding language that has allowed for accomplishing the transfer of knowledge in a precise way.

Por y para mi padre, que como en todo,

estará presente en cada página.

Índice

Resumen i
Abstract iii
Índice vii
Preámbulo 1
Capítulo 1 3
Introducción 3
Capítulo 2 13
Marco teórico 13
2.1 La Lingüística Computacional 13
2.1.1 Las aplicaciones 17
2.1.1.1 Tecnologías del texto 18
2.2 La Lingüística de Corpus 20
2.2.1 Los requisitos 23
2.2.2 Tipos de corpus 24
2.2.3 Corpus para el PLN 26
2.3 La convergencia de disciplinas 28
2.4 ¿Dónde se inscribe esta propuesta? 30

Capítulo 3 33
Estado de la cuestión 33
3.1 Aproximaciones y modelos en el PLN 34
3.1.1 Modelo simbólico 35
3.1.1.1 Las gramáticas computacionales 35
3.1.2 Modelo estadístico 37
3.1.2.1 N-gramas 38
3.1.3 Modelo conexionista 39
3.2 Cronología de las aproximaciones del PLN 40
3.3 Sistemas de verificación textual; los correctores 42
3.3.1 Dos aproximaciones para abordar el problema 42
3.3.1.1 Técnicas de bajo nivel 43
3.3.1.2 Técnicas de alto nivel 45

3.3.2 Y un problema más; la ambigüedad ... 46
3.4 Correctores ortográficos ... 47
3.5 Correctores gramaticales ... 50
3.6 Correctores de estilo ... 52
3.7 Algunos correctores desarrollados para el español ... 53
3.7.1 Un sistema basado en la identificación de patrones ... 56

Capítulo 4 ... 61
Recursos lingüísticos para el PLN ... 61
4.1 Las bases de datos léxicas ... 62
4.1.1 Teoría de la anotación ... 65
4.2 Los corpus ... 68
4.2.1 La codificación ... 71
4.2.2 La anotación ... 71
4.2.3 Corpus TIP ... 73
4.3 Las herramientas de análisis lingüístico ... 74
4.3.1 Un lematizador ... 76
4.3.2 Un flexionador ... 77
4.3.3 Un conjugador ... 78

Capítulo 5 ... 81
Propuesta de etiquetario para el español. El Lexicón TIP ... 81
5.1 Contenido ... 81
5.2 Fuentes de extracción y su tratamiento ... 82
5.3 Cobertura. El lexicón en cifras ... 84
5.4 Información compilada sobre los componentes ... 85
5.5 Concreción de la propuesta ... 87
5.5.1 Fundamentación lingüística ... 88
5.5.1.1 Preámbulo. Sobre las unidades de la gramática ... 88
5.5.1.2 Perspectivas gramaticales ... 91
5.5.1.3 Categorías léxicas y categorías funcionales ... 92
5.5.1.4 Las categorías funcionales del etiquetario ... 94
5.5.1.5 Criterios para la identificación de unidades. Límites y fortalezas ... 95
5.5.1.6 Flexión y sintaxis ... 99

5.6 Los determinantes 101
5.6.1 Consideraciones previas 102
5.6.2 Las formas potenciales: problemas relacionados 103
5.6.3 Los adjetivos determinativos 104
5.6.4 Adjetivos y pronombres: la duplicación de paradigmas 106
5.6.5 Gramática de los determinantes 107
5.6.6 Los elementos del paradigma 109
5.7 Las clases transversales 111

Capítulo 6 113
PatErr: estrategias para su constitución 113
6.1 Planteamientos generales 113
6.2 Concreciones 115
6.3 Metodología 121
6.3.1 Selección del contenido de los patrones 121
6.3.2 Estudio del fenómeno y acotación de la casuística 122
6.3.3 La viabilidad de los patrones: formalización y codificación 123
6.3.4 Tratamiento del error 125
6.3.5 Anotación de otros rasgos 127
6.4 Formalización del error: los patrones 127
6.4.1 Patrones codificados 129
6.4.1.1 Orden de prelación de los patrones 131

Capítulo 7 133
Formalización de los patrones de error. Lenguaje de codificación 133
7.1 Las marcas de codificación 134
7.1.1 Las llaves «{ }» 136
7.1.2 La almohadilla «#» 137
7.1.3 La arroba «@» 137
7.1.4 El acento circunflejo «^» 138
7.1.5 El interrogante de cierre «?» 138
7.1.6 El signo de adición «+» 139
7.1.7 La barra invertida con una letra «\w» 140
7.1.8 Los corchetes «[ ]», el signo menos «-» y el signo más «+» 141

7.1.9 La barra vertical «|» 142
7.1.10 Consideraciones 142
7.1.10.1 Múltiples opciones para codificar un mismo error 142
7.1.10.2 Las mayúsculas 145
7.1.10.3 La cadena de búsqueda en la secuencia de tratamiento 146
7.1.10.4 Códigos en la solución 149
7.1.10.5 Generación de formas erróneas 151

Capítulo 8 153
Tipologías de error para sistemas de corrección automática 153
8.1 Panorámica 154
8.1.1 Veronis (1988) 154
8.1.2 Corder (1991) 155
8.1.3 Ramírez Bustamante *et al.* (1994) 155
8.1.4 Oliva (1997) 156
8.1.5 Verberne (2002) 156
8.1.6 Díaz Villa (2005) 156
8.1.7 Wedbjer Rambell (1999-2000) 157
8.1.8 Proyecto *GramCheck* (1996) 157
8.1.9 *CON-TEXT* (1998) 158
8.1.10 *Stilus* (2002) 159
8.2 Delimitación del concepto *error* 160
8.3 Propuesta de tipología de errores 161
8.3.1 Grado de severidad 162
8.3.2 Niveles lingüísticos 163

Capítulo 9 167
PatErr: esquema de anotación 167
9.1 Alojamiento 168
9.2 Estructura 169
9.2.1 Error 169
9.2.2 Patrón 170
9.2.3 Error general 171
9.2.4 Acción correctiva 172

9.2.5 Nivel lingüístico 173
9.2.6 Incidencia 174
9.2.6.1 Errores 174
9.2.6.2 Recomendaciones 175
9.2.6.3 Avisos lingüísticos 175
9.2.7 Fenómeno asociado 176
9.2.8 Registro lingüístico 176
9.2.9 Variedad lingüística 177
9.2.10 Contexto 177
9.2.11 Cotexto 178
9.2.12 Condición 179
9.2.13 Lista asociada 179
9.2.14 Relaciones morfoléxicas 180
9.2.15 Referencia normativa 180
9.2.16 Fuente de extracción 181
9.2.17 Nota didáctica 182
9.2.18 Consideraciones internas 183
9.3 Metodología derivada 184

Capítulo 10 185
PatErr: glosario de los errores tratados 185
Capítulo 11 189
Ortografía 189
11.1 Homófonos 189
11.1.1 Avisos lingüísticos: algunos pares habituales 190
11.1.2 Pares frecuentes con tratamiento automático 192
11.1.2.1 a, ha 193
11.1.2.2 e, he 193
11.1.2.3 iba, iva 193
11.2 *hecho* y *echo* 194
11.2.1 Contextos de las formas participiales: *hecho, hecha, hechos, hechas* 195
11.2.2 Contextos del sustantivo *hecho* 195
11.2.3 Contextos de la locución *de hecho* 195
11.2.4 Contextos de las flexiones de *echar*: *echo, echa* y *echas* 195

11.2.5 Errores que involucran formas participiales de *hacer* 197
11.2.6 Errores que involucran formas flexionadas de *echar* 201
11.2.7 Errores que involucran al sustantivo *hecho* 202
11.2.8 Errores que involucran la locución *de hecho* 203
11.3 deshecho y desecho 203
11.3.1 Contextos de las formas participiales de *deshacer* 204
11.3.2 Contextos de las formas flexionadas de *desechar* 204
11.3.3 Contextos del sustantivo *desecho* 205
11.3.4 Errores que involucran formas participiales de *deshacer* 205
11.3.5 Errores que involucran formas flexionadas de *desechar* 207
11.3.6 Errores que involucran al sustantivo *desecho* 209
11.4 *ahí*, *ay*, *hay* 209
11.4.1 Errores que involucran el adverbio *ahí* 210
11.4.2 Errores que involucran el verbo *hay* 210
11.4.3 Aviso lingüístico 211
11.5 Casi homófonos 211
11.5.1 Errores 212
11.5.2 Avisos lingüísticos 213
11.6 Ortografía de algunas palabras 215
11.7 Palabras que admiten dos grafías 216

Capítulo 12 219
Morfología 219
12.1 El plural de algunos sustantivos compuestos 219
12.2 *Pluralia tantum* 221
12.3 *Singularia tantum* 222
12.4 Sustantivos terminados en *-í, -ú* 222
12.5 Sustantivos terminados en *-y* 223
12.6 Sustantivos sin flexión de plural 224
12.7 Algunos monosílabos 226
12.8 Algunas expresiones compuestas 227

Capítulo 13 229

Gramática 229
13.1 Concordancia 230
13.1.1 Concordancia intrasintagmática; categorías 230
13.1.1.1 La forma poco 231
13.1.1.2 Cuantificadores ambiguos 232
13.1.2 Concordancia intersintagmática; relaciones entre constituyentes oracionales 234
13.1.2.1 Sustantivos clasificativos 235
13.1.2.2 Sustantivos colectivos 235
13.1.2.3 Pronombres indefinidos 236
13.1.2.4 El verbo faltar 237
13.1.2.5 Estructuras copulativas 238
13.1.2.6 Algunas expresiones 239
13.1.2.7 Construcciones partitivas 241
13.1.2.8 Oraciones escindidas 242
13.1.3 Concordancias de género 246
13.1.3.1 Vacilación en el género de algunos sustantivos 246
13.1.3.2 Profesiones u oficios 247
13.1.3.3 Cuantificadores e indefinidos 248
13.1.3.4 Algunos sintagmas preposicionales 249
13.1.3.5 Tratamientos de respeto 249
13.1.3.6 Género de algunos numerales 250
13.1.3.7 Excepciones; algunos sustantivos femeninos 251
13.1.4 Concordancias de número 253
13.1.5 Concordancias de persona 255
13.2 Gramática verbal 257
13.2.1 Usos del gerundio 258
13.2.2 Usos del infinitivo 260
13.2.3 Vacilación entre tiempos o modos verbales 262
13.2.3.1 Usos no impersonales de verbos que lo son 263
13.2.3.2 Vacilación de tiempos verbales 268
13.3 Dequeísmo 269
13.3.1 Oraciones subordinadas sustantivas 269
13.3.1.1 Algunas expresiones copulativas 273
13.3.1.2 Oraciones subordinadas sustantivas con función de atributo 274

13.3.1.3 Verbos con vacilación en su rección preposicional ... 274
13.3.2 Locuciones conjuntivas ... 275
13.4 Queísmo ... 275
13.4.1 Supresión de la preposición *de* ... 276
13.4.1.1 Queísmo en régimen verbal general ... 276
13.4.1.2 Nexos subordinantes y locuciones conjuntivas ... 277
13.4.1.3 Expresiones o perífrasis verbales ... 277
13.4.1.4 Expresiones con sustantivos ... 278
13.4.1.5 Expresiones con adjetivos ... 278
13.4.2 Supresión indebida de otras preposiciones ... 279
13.5 Régimen preposicional ... 279
13.5.1 Errores de régimen preposicional en términos o expresiones ... 279
*13.5.1.1 *adicción con, *opción a, *idéntico con ... 279*
*13.5.1.2 *mayor a, *mayor de ... 280*
*13.5.1.3 *a la mayor brevedad posible ... 281*
13.5.2 Errores de régimen preposicional en verbos ... 281
13.5.3 Errores con verbos que rechazan preposición ... 281
13.5.4 Avisos de régimen preposicional en verbos ... 282

Capítulo 14 ... 283
Léxico ... 283
14.1 Impropiedades léxicas ... 283
14.1.1 *Cesar*, *dimitir* y *destituir* ... 284
14.1.2 *Cuyo* despojado de posesividad ... 285
14.1.3 Restricciones semánticas e impropiedades en algunos verbos ... 286
14.2 Precisión ... 287
14.3 Neologismos ... 288
14.4 Redundancia ... 289
14.4.1 Estructuras genéricas ... 290
14.4.2 Expresiones redundantes frecuentes ... 292

Capítulo 15 ... 295
Estilo ... 295
15.1 Dos formas para decir lo mismo ... 295

15.1.1 Palabras 296
15.1.1.1 Superlativos 296
15.1.1.2 Miscelánea 297
15.1.2 Expresiones 298
15.1.2.1 Expresiones partitivas 298
15.1.2.2 Expresiones con dar 300
15.1.2.3 Expresiones reflexivas 300
15.1.2.4 Expresiones muy frecuentes 301
15.2 Coloquialismos 302
15.2.1 Inicio de frase 302
15.2.2 *de seguido* 303
15.2.3 *Más* intensificador 304
15.2.4 Expresión de la duda 304

Capítulo 16 305
Tratamientos transversales 305
16.1 Construcción de una expresión 305
16.1.1 Errores en la ortografía 306
16.1.1.1 Adverbios prefijados con a- 306
16.1.2 Errores gramaticales 307
16.1.2.1 Rección de algunos adverbios de lugar 307
16.1.2.2 De impropio en lugar de la conjunción que 308
16.1.2.3 Algunas expresiones recíprocas 308
16.1.2.4 Errores causados por la rección preposicional 309
16.1.2.5 Algunos pronombres indefinidos 313
16.1.3 Errores de estilo 314
16.1.3.1 Expresiones de tiempo: adverbios y preposiciones 314
16.1.3.2 Expresiones numerales 315
16.1.3.3 Expresiones de enfoque 316
16.1.3.4 Cacofonías 317
16.1.4 Incidencias en el plano léxico-semántico 317
16.1.4.1 Interferencias entre construcciones 318
16.1.4.2 Restricciones semánticas 318
16.2 Extranjerismos 320
16.2.1 Léxico: extranjerismos con versión española 321

16.2.2 Ortografía: extranjerismos adaptables al español 325
16.2.3 Tipografía: extranjerismos y locuciones latinas 326
16.2.4 Estilo: expresiones tomadas de estructuras foráneas 327
16.3 Expresiones y términos latinos 330
16.3.1 Ortografía y expresión correcta 330
16.3.2 Morfología: el plural 331
16.3.3 La tipografía 334
16.4 Segmentación y unificación 334
16.4.1 Sin cambio de significado 335
16.4.1.1 Términos para los que se recomienda la grafía univerbal 335
16.4.1.2 Términos para los que se recomienda la grafía segmentada 336
16.4.2 Con cambio de significado 337
16.4.3 *Por qué*; *porque*; *por que* 339
16.4.3.1 Por qué 339
16.4.3.2 Porque 341
16.4.3.3 Porqué 342
16.4.3.4 El aviso lingüístico 343
16.4.4 *Sino*; *si no* 344
16.5 Vulgarismos 347

Conclusiones y horizontes futuros 349
Conclusiones 349
Trabajos futuros 351

Referencias 355
Anexo 1. Cuantificadores 365
Anexo 2. Numerales 367
Anexo 3. Indefinidos 379
Anexo 4. Posesivos 381
Anexo 5. Relativos 383
Anexo 6. Interrogativos y exclamativos 385
Anexo 7. Demostrativos 387
Anexo 8. Repertorio de listas 389
Anexo 9. Abreviaturas 391
Anexo 10. Incidencias 393
Anexo 11. Fenómenos asociados 401

Anexo 12. Zonas geográficas 405
Anexo 13. Adverbios de grado 407
Anexo 14. Adjetivos de grado extremo 409
Anexo 15. Locuciones latinas 411

Preámbulo

Pocos conceptos hay que sean tan importantes para la humanidad como lo es el lenguaje. Cuando se trata de describir al ser humano, ya sea en términos físicos o filosóficos, casi siempre aparece como la característica diferencial más clara.

Los científicos no podían permanecer ajenos a una llamada tan intensa, pero las cualidades particulares del lenguaje han hecho que se haya tardado siglos en encontrarle un hueco propio dentro de las investigaciones. De hecho, no es hasta finales del siglo XIX cuando se sitúa el origen de la lingüística contemporánea. Este retraso, no obstante, no evita que se hayan acumulado años de investigación esforzada y una imponente producción de textos valiosísimos que explican hasta el más recóndito aspecto del lenguaje y sus lenguas.

Como ciencia, la lingüística intenta aproximarse a su objeto de la misma forma en que lo hacen las ciencias de la naturaleza, ateniéndose a cánones propios de la investigación científica; objetividad, coherencia, simplicidad, etc. Por tratar con los datos del lenguaje se dice que es empírica, y como otras ciencias de esta naturaleza, su investigación científica explora, describe, explica y predice los acontecimientos confrontando las aseveraciones de sus enunciados con una base empírica de datos. Pero sus límites, como los de la mayoría de las disciplinas, son flexibles, más aún cuando su objeto es infinito.

La lingüística no ha podido permanecer indolente a la revolución tecnológica. De hecho, parece que esta capacidad que nos define, junto con la utilización y desarrollo de tecnologías complejas constituyen dos de las características esenciales del hombre. Parece indudable que el desarrollo tecnológico depende del lenguaje; resulta difícil concebir las tareas de ingeniería —desde el desarrollo del cincel hasta el de un acelerador de partículas—, exentas de comunicación, de lengua.

El espectacular desarrollo de la ingeniería de los ordenadores durante la segunda mitad del siglo XX ha permitido la utilización de procedimientos y métodos informáticos para el tratamiento automático del lenguaje. Esta nueva

realidad, casi inexplorada por la lingüística general, aporta a nuestra disciplina perspectivas inéditas y posibilidades antes insondables, que abren múltiples espacios aún sin franquear y recuerdan que otros, más tradicionales, necesitan ser revisitados. La lingüística, en la actualidad, ha dejado de ser un fin en sí mismo para ser un medio muy productivo que colabora ostensiblemente en la mejora de la realidad virtual y tecnológica del mismo modo que esta se presta a facilitar nuestra expresión lingüística traspasando las barreras que hace no tanto tiempo limitaban la comunicación.

Inserta en esta dialéctica, nuestra ciencia se presenta en constante desarrollo, descubriendo remotos horizontes e incorporando nuevos espacios a la investigación, los cuales desbordan los límites del dominio que ella misma se marcó cuando definió su objeto al constituirse como ciencia.

Capítulo 1

Introducción

Las llamadas INDUSTRIAS DE LA LENGUA investigan y desarrollan una amplia variedad de herramientas que se aplican sobre el texto escrito. Actualmente, buena parte de los esfuerzos investigativos y económicos se centran en el campo de la revisión textual, ya sea parcial o global. Para este fin, se diseñan y articulan programas capaces de verificar, cada vez con mayor eficacia, cobertura y precisión lingüística, la corrección textual.

Posiblemente esta incorporación en el mercado haya sido la causa que ha concitado la atención de parte de la investigación lingüística reciente, pero es esta misma circunstancia la que ha motivado que los resultados de estos trabajos y los logros que se han conquistado haya que deducirlos a partir de la observación del producto final, sin tener acceso ni al proceso y ni a la materia que los ha gestado. Poco hay escrito, con cierto grado de profundidad, sobre los planteamientos, procedimientos y metodologías que se siguen para el desarrollo de estos programas. La bibliografía concreta que aborda el asunto de los correctores para el español es escasa, parcial, y se agota en la superficie, aunque en algunos casos resulta sugerente.

El desarrollo de los correctores automáticos ha estado vinculado a la búsqueda e implementación de técnicas especiales para la detección e intervención sobre las secuencias erróneas. Los correctores, como casi todas las aplicaciones que se desarrollan en el área del Procesamiento del Lenguaje Natural, dependen, entre otras circunstancias, del dominio donde se vayan a aplicar, de la información lingüística sobre la que se fundamenten y de la lengua objeto del tratamiento.

Por otro lado, los actuales programas de verificación y corrección ortográfica, gramatical y de estilo suelen encontrarse incorporados en la mayoría de los

procesadores de textos, aunque también se ofrecen como una funcionalidad adicional en aplicaciones de diversa naturaleza en las que se hace necesaria la comunicación textual.

Señala Gómez Guinovart en un artículo general sobre Lingüística Computacional que la opción más invocada para llevar a cabo la corrección gramatical es la siguiente:

> A técnica informática máis utilizada para a identificación dos erros gramaticais ten un enfoque casuístico e baséase no recoñecemento de certos patróns de erro previamente establecidos (2000:18)

Este planteamiento permite sustituir, para el caso de un corrector gramatical, el análisis sintáctico de un texto en busca de irregularidades estructurales, por la identificación de secuencias erróneas recogidas en patrones de error. El análisis sintáctico automatizado exige la formalización de reglas que atrapen el conocimiento lingüístico que deriva de una gramática. Pero, para su aplicación en un corrector estas reglas formalizadas, que constituyen una gramática computacional, deberán ser capaces de analizar tanto lo estructuralmente correcto como los desvíos que con frecuencia se registran en la lengua.

El recurso que aquí presentamos, PatErr, se actualiza en un nutrido repertorio de patrones de error codificados del español panhispánico que se inspira en los planteamientos de la citada técnica de identificación de patrones y la asume no solo para la corrección gramatical, sino para una corrección global, que aborde todos los niveles de la lengua. Este compendio de patrones, aliado con algunas técnicas básicas de análisis lingüístico automático, constituye un recurso competitivo en el marco de lo que hemos denominado industrias de lengua.

La consecuencia inmediata que la aparta de otros posibles desarrollos será la sustitución de reglas gramaticales por una batería de datos formalizados que contienen errores, esto es, malformaciones estructurales e incorrecciones registradas en los diversos niveles de la lengua. El objeto de la formalización

lingüística y, por lo tanto, del enfoque de este planteamiento será, pues, un producto de la actuación[1], y no materia abstracta arrojada por la competencia.

El error, por lo tanto, debe ser el punto de partida que motive el diseño y el desarrollo de cualquier recurso automático de revisión, corrección y asesoría lingüística. Será este el objeto de estudio y formalización computacional y permitirá, inserto en una suerte de disciplina que denominaremos *lingüística de errores con fines computacionales,* el estudio científico del lenguaje partiendo de una muestra parcial de sus infinitas producciones; el error. De este se podrán extraer inferencias e hipótesis que permitan modelizar y formalizar el lenguaje a partir de los datos lingüísticos que presentan violaciones con respecto a lo gramatical.

Para que este planteamiento pueda ofrecer las garantías y cobertura exigibles en el entorno de las industrias de la lengua, estos patrones de error deben, idealmente, captar toda la casuística de desvíos o incorrecciones que puedan surgir en el uso de la lengua escrita. Para esto, parece necesario llevar a cabo un estudio exhaustivo tanto empírico como teórico sobre la *casuística de lo agramatical* con el objetivo de registrar el resultado de estas investigaciones en forma de patrones de error formalizados que permitan la identificación de errores en el texto por parte de un mecanismo autómata.

Aunque, como detectaba Gómez Guinovart el procedimiento de identificación de errores parece ser la técnica informática más utilizada para el desarrollo de los correctores gramaticales, no se han encontrado, como se anunció, publicaciones, documentos o especificaciones sobre este tipo de programas en los que se expongan las bases o criterios que describan o aporten información concreta sobre cómo llevar a término la tarea de diseño y desarrollo de un programa de revisión lingüística basada en la técnica de identificación de patrones.

Esta ausencia de publicaciones hace imposible, entre otras cosas, la comparación de este trabajo con otros sistemas similares[2]. Por este motivo aunque

[1] Se propone, con toda la intención, la expresión *un producto de la actuación* porque entendemos que el error es uno de los tipos de producto —*defectuoso*— que deriva del uso de la lengua.

[2] Este hecho, la ausencia de publicaciones expositivas o valorativas relacionadas con el desarrollo de *software*, ha sido una constante a lo largo del desarrollo de este trabajo. Las escasas publicaciones dentro de este ámbito no ofrecen concreciones ni desvelan las entrañas de los programas, ni los procesos o procedimientos que han sido utilizados para su desarrollo. Esta realidad posiblemente surja

el enfoque que aquí adoptamos sea el *más utilizado,* las labores de ingeniería lingüística y su concreciones —metodología, recursos necesarios y desarrollo— que aquí se proponen explícitamente, son originales y propias, en el sentido de que emergen a partir de los recursos, necesidades, límites y potencias del entorno en el que surge[3].

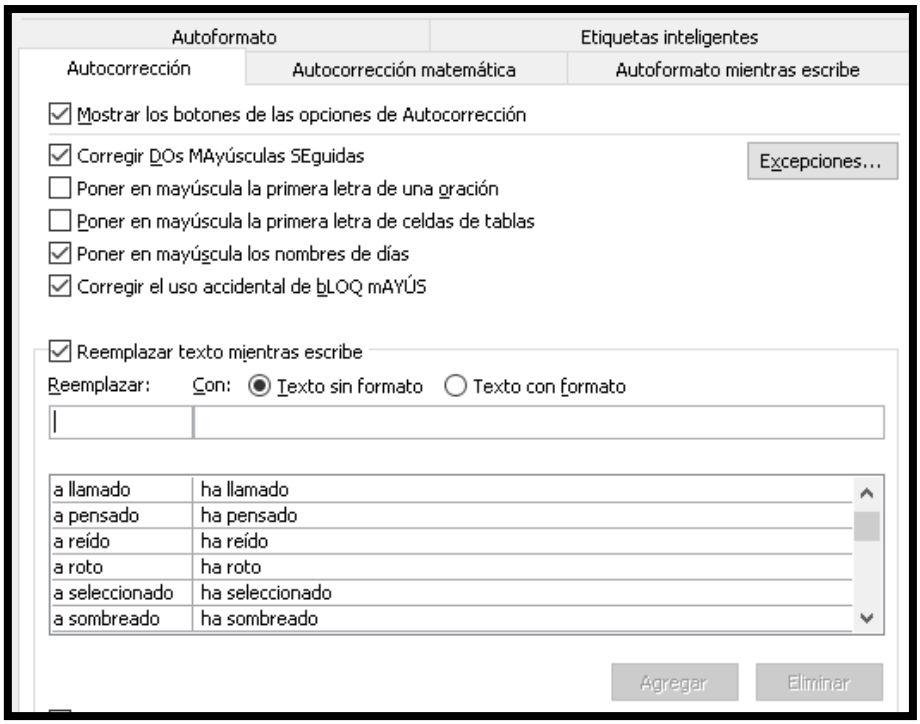
El desarrollo de esta propuesta, entonces, deberá apoyarse en exploraciones empíricas sobre otros programas que parecen asentarse, en algunos de sus procesos, en esta técnica. Este parece ser el caso del corrector automático alojado en el procesador de textos de *Microsoft Word* 2016. A partir del análisis de las pruebas que se han llevado a cabo sobre este programa pueden ilustrarse algunas de sus limitaciones e inconsistencias que esta propuesta pretende subsanar.

Se ha observado que, mientras este corrector es capaz de ejecutar una corrección para la secuencia **a comido*, resulta incapaz de reconocer el error y tratarlo en secuencias similares como **a resultado, *a tenido, *a vertebrado,* o **a colegido*. Un primer análisis de estos resultados revela que el programa adolece de inconsistencia en términos de cobertura. Está limitado para ofrecer un tratamiento a toda la casuística que deriva de un mismo fenómeno, pudiendo solo asistir a unos cuantos casos aislados.

En la siguiente imagen se observan seis de los 18 patrones desarrollados para solventar los errores asociados a este fenómeno.

ante el horizonte mercantil que finalmente aflora en este tipo de productos desarrollados en el marco del Procesamiento del Lenguaje Natural.

[3] Este trabajo surge al amparo del grupo de investigación CLTIP, adscrito al *Instituto universitario de análisis y aplicaciones textuales* IATEXT, de la Universidad de Las Palmas de Gran Canaria.



Por otro lado, la presencia de los patrones tampoco parece obedecer a un criterio de frecuencia de la palabra. De otro modo, la secuencia errónea intratada **a tenido* debería ser identificada con preferencia a la identificación de la secuencia **a sombreado.*

Los patrones de error que aquí proponemos, sospechamos que a diferencia de los patrones que operan en otros programas de corrección, presentan una naturaleza diferente. Su contenido, además de formalizado está codificado, lo que permite, en virtud de estas marcas de codificación, crear abstracciones y generalizaciones que trasciendan la literalidad de los patrones y atrapen todas las formas posibles que pueda adoptar un mismo error. Se pretende, pues, optimizar la cobertura de cada patrón compilado mediante marcas de codificación que, inspiradas en las expresiones regulares e interpretadas por un autómata, posibilitarán al menos dos cauces para llevar a cabo la corrección. Por un lado, las marcas, asociadas a un *sistema generativo*, servirán para provocar una generación automática de patrones de error —exentos de marcas de codificación— capaces de copar toda la casuística que derive de cada fenómeno o estructura errónea. Por otro, estos patrones podrán contrastarse directamente con el texto susceptible de corrección, a tiempo real, siempre y cuando se disponga de un *sistema de etiquetado* que actúe previamente sobre el texto.

El tipo de patrones, junto con su tratamiento, que para un error como el expuesto se desarrollará en el marco de esta propuesta tiene la siguiente estructura:

Patrón de error	Corrección
a + participio	***ha*** + participio

Diseñado así, tendrá la capacidad de corregir cualquier secuencia que presente esta configuración en el texto debido a que este patrón puede dar lugar, a partir de herramientas de descodificación, a un conjunto de patrones derivados —uno para cada participio que recoja el lexicón sobre el que se asiente— que serán los que finalmente deban buscarse e identificarse en el texto.

Por otro lado, si se opta por la opción de ejecución a tiempo real sobre un texto etiquetado, la estructura abstracta podrá identificar cualquier réplica errónea que se concrete en el texto. De este modo, con independencia del modo en que se aplique el patrón, se puede garantizar un tratamiento unitario para toda la casuística errónea agazapada entre los múltiples casos que puede acoger el mismo fenómeno.

Como se observa, el recurso que aquí proponemos pretende llevar a cabo un tratamiento global de algunos fenómenos que subsumen cientos de errores en potencia. La generación automática de toda la casuística o la identificación de errores a partir de estructuras más abstractas garantiza así, tanto la corrección de secuencias frecuentes **a comido*, como la de expresiones más residuales como **a craneado, *a frutecido* o **a añusgado.* Este afán globalizador, que será una constante a lo largo de todo este trabajo, se postula con el único fin de evitar inconsistencias en términos de cobertura.

La puesta en marcha de un sistema de revisión textual que se nutra de un depósito de patrones de error como el que aquí planteamos requiere, en primera instancia, la formalización en una base de datos de patrones de error codificados capaces de captar toda la casuística de errores asociados a un fenómeno. El diseño y desarrollo de este compendio de patrones —PatErr—, que se constituye para

ofrecerse como un recurso robusto dentro del ámbito de la revisión textual, será, como se verá, el fundamento de esta tesis.

Para la consecución de estos dos objetivos —diseño y desarrollo de PatErr—, ha sido necesario disponer de ciertos componentes que constituyen un sistema de actuación unitaria que, bajo el mandato de unas marcas de codificación, —un lenguaje común— son capaces de generar masivamente los errores. Esta creación artificial se basará en la explotación de la forma o formas codificadas en el patrón, bien mediante la flexión o la conjugación, bien a partir del desarrollo léxico de una categoría gramatical, bien a través de listas de términos que el patrón se arroga para dar cabida a todas las opciones posibles de error.

La productividad tanto de las marcas de codificación, como de los recursos que las interpretan —lematizador, flexionador y conjugador TIP— es indiscutible debido a que permiten desarrollar o identificar patrones sincréticos de gran alcance que garantizan la consistencia del programa autómata que lo explote.

Como es obvio, previa a la formalización y codificación de los patrones, es necesario un estudio de corte empírico y teórico de la casuística que confirme la viabilidad y la universalidad del patrón codificado con el fin de evitar falsos positivos y correcciones indebidas. Para llevar a cabo estas tareas propias de lo que hemos denominado lingüística de errores con fines computacionales ha sido indispensable otro recurso esencial en el entorno de estos trabajos de Procesamiento del Lenguaje Natural; un corpus de frases del español escrito, Corpus TIP, que ha servido como fuente inagotable de datos sobre el uso de la lengua en el español escrito.

Por otro lado, la cobertura que pueda ofrecer un repertorio de errores depende, en gran medida, del lexicón sobre el que descanse. En el caso de PatErr se ha utilizado el Lexicón TIP, sobre el que se ha llevado a cabo un trabajo previo que puede sistematizarse en tres líneas de acción; revisión, corrección y ampliación de la información registrada a partir de procedimientos de extracción automáticos ejecutados sobre diversas fuentes lexicográficas.

El horizonte de referencia, presente en todo el trabajo lingüístico que se ha llevado a cabo, será el español normativo recogido tanto en la *Nueva Gramática de la lengua española*, como en otros documentos elaborados por la Academia[4].

Como se colige de esta breve descripción, para el desarrollo y explotación de un recurso como PatErr que se ofrece para la revisión textual automática basada en la identificación de patrones de error, se hace necesaria la presencia de dos líneas de trabajo; una relacionada con el *conocimiento lingüístico*, su análisis y formalización en patrones y reglas computacionalmente comprensibles para la ingeniería informática, y otra relacionada con el *tratamiento computacional* de la información lingüística y el diseño y la creación de aplicaciones finales para el usuario.

A lo largo de estas páginas se abordarán todos los aspectos relacionados con la primera línea de trabajo, la lingüística. Se presentarán las características del repositorio de errores PatErr, sus rasgos de diseño, su estructura y la naturaleza de la información que contiene, el lexicón que le ha servido de base, las herramientas y recursos que han asistido a la investigación lingüística, el lenguaje de codificación propio que se ha desarrollado y otras tantas especificaciones relevantes que ayuden a esbozar por un lado, la naturaleza de los tareas que surgen en el entorno de la lingüística de errores con fines computacionales, y por otro, el potencial que el repertorio de errores codificados que presentamos puede alcanzar actualizado en un programa de ayuda a la escritura del español que ofrecerá servicios de corrección y asesoría lingüística.

Debe advertirse, no obstante, que la utilidad de este recurso, en el punto de desarrollo en el que se encuentra, está sujeta a unos mínimos de legibilidad que debe presentar el texto revisable. La ausencia de un módulo que ofrezca tratamiento ortográfico global o de un motor que revise las concordancias —previsto para desarrollar en el futuro como una pieza separada— exige que el texto que se someta al programa que explote el contenido de PatErr ofrezca unas mínimas garantías de calidad y demuestre cierto grado de competencia en la gramática española. Aunque su continuo desarrollo podrá derivar en un programa

4 No obstante, para el desarrollo de los patrones relacionados con la ortotipografía, se seguirán, principalmente, los criterios de los tratados de Martínez de Sousa (2000, 2004).

autónomo, en la actualidad su explotación debe orientarse a la creación de un programa complementario para los correctores alojados en los procesadores de texto.

No obstante, su viabilidad en términos computacionales, se demuestra en un prototipo desarrollado sobre un estado embrionario de este trabajo que recibió un premio en la convocatoria *Talento y Compromiso* que llevó a cabo Cajasiete en el año 2015[5].

Por último, cabe advertir que, si el lenguaje es infinito, los errores que con su materia puedan surgir son aún más inescrutables. Conscientes de esta limitación que emerge de la propia naturaleza del objeto que tratamos, este trabajo enfocado desde la *lingüística de errores con fines computacionales* surge con el ánimo de ofrecer, a partir de la reutilización de recursos básicos para el procesamiento del lenguaje natural, unas bases y un contenido sobre los que desarrollar una herramienta de ayuda a la escritura capaz de integrarse competitivamente en las industrias de la legua española.

[5] http://tip.iatext.ulpgc.es/ASLI/Corrector/MenuCorrector

Capítulo 2

Marco teórico

El marco teórico en el que se inscribe el trabajo que pretende describir esta tesis se construye a partir de la convergencia, cada vez más recurrente, de dos disciplinas o enfoques que se dedican al análisis de las lenguas. Estas son la Lingüística Computacional y la Lingüística de Corpus que tienen distintos rigores y finalidades, pero logros compartidos.
No como marco o recipiente, sino como el contenido sobre el que se obra, participa la gramática española, que desde la intuición hasta los principios, desde el uso hasta la norma, desde la excepción hasta la regla, da materia a toda la propuesta. A continuación, se ofrece una panorámica de ambas perspectivas que pretende surtir de los principios, claves y contextos necesarios para circunscribir la investigación en su coordenada precisa.

2.1 La Lingüística Computacional

Uno de los axiomas de la lingüística moderna sostiene que las lenguas son mecanismos complejos que relacionan secuencias de *significantes* con conceptos o *significados*. La tarea del lingüista consiste en analizarlas con el fin de descubrir principios generales que expliquen su funcionamiento. Pero la LINGÜÍSTICA COMPUTACIONAL, que promueve el estudio del lenguaje desde una nueva perspectiva, no se detiene en el análisis; su objetivo es el tratamiento automatizado de estas secuencias y el dominio del sistema de la lengua por parte de la máquina.

La LINGÜÍSTICA COMPUTACIONAL surge, pues, como una parcela de conocimiento interdisciplinar que emerge de disciplinas como la lingüística, la informática, la ciencia cognitiva o la ingeniería electrónica. Su búsqueda se orienta al hallazgo de métodos que incorporen en las computadoras habilidades para el manejo y tratamiento informatizado de las lenguas y la información. Es una parte esencial de la Inteligencia Artificial, que investiga y desarrolla mecanismos

computacionalmente efectivos capaces de analizar, entender y generar textos, tanto orales como escritos, basados en una lengua natural (Moreno *et al*., 1999; Gómez, Guinovart, 2000)[6].

Bajo la denominación de Lingüística Computacional convive un conjunto relativamente heterogéneo de teorías, métodos, herramientas, aplicaciones y productos que tienen en común la consideración de la lengua como un objeto susceptible de ser tratado mediante procedimientos informáticos. De esta premisa se derivan tres líneas de actuación; (i) la elaboración de modelos lingüísticos en términos formales e implementables, (ii) la aplicación de estos modelos a cualquiera de los niveles de descripción lingüística y, (iii) la comprobación automatizada de la congruencia de una teoría lingüística y sus predicciones (Gómez Guinovart, 2002:221). Ante un campo de acción tan amplio y diverso, surgen diferentes necesidades y objetivos que se saciarán con soluciones de diferente naturaleza y arquitectura.

Habida cuenta del galimatías terminológico que suscita la interrelación entre la lingüística y la informática —Lingüística Computacional, Procesamiento del Lenguaje Natural, Ingeniería Lingüística, Lingüística Informática, Tecnologías del lenguaje, Industrias de la lengua, etc.— conviene hacer una somera topografía que ayude a delimitar cada ámbito junto con sus particularidades y objetivos[7].

A pesar de los matices y postillas que surgen de cada definición, con el término de LINGÜÍSTICA INFORMÁTICA se suele hacer referencia al uso de herramientas informáticas en la investigación lingüística. Es, pues, una disciplina que se orienta al desarrollo de programas de apoyo al trabajo filológico, lexicográfico, lingüístico o humanístico[8].

[6] Se entiende, a grandes trazos, que la Inteligencia Artificial se dedica al desarrollo de sistemas informáticos que simulan el comportamiento humano. Algunas de sus ramas investigan en parcelas referidas a la adquisición de conocimientos, al razonamiento que puede surgir a partir de estos o la representación de esos conocimientos adquiridos.

[7] Para el desarrollo de este esbozo se ha recurrido a las propuestas generales de Moure y Llisterri (1996), Rojo (2002), Ruíz Antón (2005), Martí y Llisterri (2002) y Llisterri (2007), Periñán (2005, 2012). Para otras clasificaciones alternativas acerca de estas áreas, puede consultarse Gómez Guinovart (2000).

[8] En estos casos, las herramientas informáticas se constituyen como elementos auxiliares válidos para la elaboración de diccionarios, la constitución de repertorios terminológicos, atlas lingüísticos, la puesta en marcha de programas que aportan información de frecuencia o distribución de elementos, etc. Manuales como los de Hockey (1980) o Butler (1985) son representativos de esta corriente (Moure y Llisterri, 1996: 148 y ss.).

Frente a la utilización instrumental de la informática, que se desarrolló sobre todo en los años 80 surge un nuevo campo, la Lingüística Computacional —en adelante LC—, en el que se integran programas informáticos de uso local, en la red o en entornos que requieren la interacción entre personas y ordenadores para permitir el tratamiento de las lenguas, sea en su vertiente oral o escrita. El campo de actuación de esta ciencia se centra tanto en el diseño de algoritmos y estructuras de datos útiles para la representación y el procesamiento de los datos lingüísticos, como en el desarrollo de programas informáticos cuya utilidad derive de utilizar conocimiento lingüístico de algún tipo.

Las líneas esenciales que caracterizan a la LC actual son, siguiendo a *Biber et al.* (1998:6), las siguientes[9]:

I. Es empírica. Analiza patrones reales de uso en contextos naturales.
II. Utiliza grandes colecciones de textos, los corpus, que sirven de base para sus análisis.
III. Hace uso de técnicas automáticas e interactivas a través de ordenadores.
IV. Se basa en técnicas analíticas y cuantitativas.

En síntesis, la LC es la disciplina que abarca tanto el procesamiento del lenguaje, como el del habla, desde una perspectiva general o desde un punto de vista teórico. La figura del lingüista, que actualmente se reconoce como esencial en cualquier investigación de esta naturaleza, deberá ayudar a transferir sus investigaciones en lingüística teórica o en cualquier otra disciplina lingüística, al entorno del modelo computacional (Periñán, 2012:30)[10].

Por su parte, la INGENIERÍA LINGÜÍSTICA, abraza un terreno que abarca las técnicas propias de la informática y se basa en la aplicación de los conocimientos lingüísticos al desarrollo de sistemas informáticos capaces de reconocer, procesar, interpretar y generar el lenguaje humano en todas sus posibilidades. En este ámbito se desarrollan tecnologías lingüísticas que hacen uso de nuestros conocimientos de la lengua para optimizar la utilización de los sistemas

[9] Para una revisión general de los límites y objetivos de la disciplina, puede consultarse Grishman (1986), Moreno Sandoval (1998), Martí y Llisterri (2002), Periñán (2005, 2012).
[10] *Anytime a linguist leaves the group the recognition rate goes up*; Jelinek, F. (cf. Periñán, 204:14).

informáticos y el acceso a las redes que configuran la sociedad de la información y del conocimiento, siendo capaces de eludir las barreras que impone la distancia, el uso de lenguas distintas o el canal en que tiene lugar la comunicación (Martí, 2003; Llisterri, 2007).

Junto a estas etiquetas, es habitual encontrar en la bibliografía relacionada la expresión INDUSTRIAS DE LA LENGUA, con la que se pretende reflejar el potencial económico y comercial del ámbito que nos ocupa. Estas industrias

> [...] abarca(n) una serie de actividades comerciales en las que el tratamiento del lenguaje, por personas, por máquinas o por una combinación de unas y otras, forma una parte fundamental del producto o servicio. De este modo, la industria de la lengua incluye métodos de publicación nuevos y tradicionales, desde los libros a las páginas de la *World Wide Web*, desde la radio a la televisión de pago. Abarca servicios ya consolidados, como la traducción, la interpretación, la transcripción y la redacción técnica, y otros servicios nuevos, como la adaptación de programas informáticos, el correo vocal y el aprendizaje de idiomas asistido por ordenador (Cdad Europea, 1997).

En este contexto, las tecnologías que se ocupan específicamente del tratamiento de la lengua oral son las llamadas TECNOLOGÍAS DEL HABLA mientras que aquellas cuyo objeto son los textos escritos se enmarcan en el PROCESAMIENTO DEL LENGUAJE NATURAL —en adelante PLN— o TECNOLOGÍAS DEL TEXTO (Llisterri, 2007).

Las tecnologías del habla tienen como objetivo el tratamiento automático de la lengua oral. Crean programas y sistemas capaces de ofrecer información hablada —*síntesis del habla*—, de reconocer enunciados emitidos por una persona, —*reconocimiento automático del habla*—, y de combinar ambas tecnologías y posibilitar la interacción entre el hombre y la máquina —*sistemas de diálogo*—[11].

El PLN o tecnologías del texto, se concretan en el tratamiento de la lengua escrita; desarrollan las herramientas necesarias para el procesamiento del lenguaje y proveen cauces para el desarrollo de tecnologías empleadas en aplicaciones que requieren sistemas tanto de generación del lenguaje, como de comprensión.

Aunque los dos campos de acción suelen quedar delimitados mediante la dicotomía *habla-lenguaje*, diferenciando a los expertos que se dedican a la señal

[11] Para una definición precisa de las áreas concretas de conocimiento y aplicación de cada uno de estos términos puede consultarse Moreno Sandoval (1998); Martí y Llisterri (2002); Martí y Taulé (2001); Periñán (2012).

sonora de los que trabajan con los textos, lo cierto es que ambas disciplinas comparten tanto herramientas, como metodología. Su desarrollo requiere el uso de recursos lingüísticos especializados como los corpus, las bases de datos léxicos, algunas herramientas de análisis lingüístico o las gramáticas computacionales y, por otro lado, los procedimientos que se utilizan para el desarrollo de estas aplicaciones hacen uso de técnicas estadísticas que parten de la información recogida en amplios corpus textuales.

2.1.1 Las aplicaciones

Actualmente existe una amplia variedad de aplicaciones y programas que tratan el lenguaje humano de alguna u otra forma. Algunas realizan un *procesamiento superficial* —aunque efectivo para ciertos fines—, que no requiere la utilización de técnicas complejas de LC; este es el caso de los correctores ortográficos que se incorporan en los procesadores de texto.

Otras investigaciones se centran en *aplicaciones de gran alcance*, que aspiran a dotar a las máquinas de capacidad para comprender el lenguaje, realizar inferencias sobre lo comprendido y participar de forma cooperativa en diálogos. Como se puede intuir, el rango de complejidad que hay entre unas y otras aplicaciones es extenso, de lo que se colige, en definitiva, que la LC se encarga de producir tecnologías y metodologías encaminadas al tratamiento de las diversas lenguas en cualquiera que sea su faceta.

En lo que se refiere a las tecnologías lingüísticas para el español, no han alcanzado aún el grado de desarrollo global que ha alcanzado el inglés o algunas otras lenguas. De acuerdo con Llisterri,

> La razón no estriba, ciertamente, en la calidad de la investigación llevada a cabo, sino en las diferencias en el peso económico entre las lenguas que dificultan que los resultados de la investigación lleguen al mercado, y por tanto, a los usuarios finales (2007:484).

Para acotar la propuesta que aquí se presenta, se ofrece una breve panorámica en la que se advierten algunos de los posibles desarrollos inscritos en las tecnologías del texto.

2.1.1.1 *Tecnologías del texto*[12]

Herramientas de ayuda a la escritura

Una de las posibles aplicaciones de las tecnologías lingüísticas son los correctores ortográficos y gramaticales, que se alojan en la mayoría de los programas de procesamiento de texto y cada vez más, en programas relacionados con la minería de datos. Podemos hacer una descripción genérica de estos programas como *herramientas de ayuda a la escritura*. Se suelen describir en este ámbito, tres niveles de complejidad creciente; la verificación ortográfica, la verificación gramatical y la verificación de estilo.

Los *correctores ortográficos* son las aplicaciones más extendidas y utilizadas por su fiabilidad aunque presentan aún ciertas limitaciones debido a que en términos generales no suelen incorporar información lingüística y se limitan a contrastar cadenas de caracteres entre el texto que se somete a la revisión y un lexicón previamente codificado. Como se puede deducir, la posibilidad de que el programa ofrezca tratamiento a las formas erróneas dependerá directamente de la cobertura del lexicón sobre el que se apoye para contrastar los errores.

Los *correctores gramaticales* pueden surtirse de una gramática computacional o bien basarse en la comparación de secuencias de palabras con unos patrones de error previamente determinados. No obstante, como señala Gómez Guinovart (2000) establecer patrones que tengan cierta validez general —evitando así el desarrollo de un patrón por cada posible error— es una operación compleja que requiere un cierto grado de abstracción lingüística.

Por último, los *verificadores de estilo* se desarrollan con el fin de detectar los elementos que no se adecúan a las orientaciones o recomendaciones que definen a la norma culta o a un estilo determinado —general, técnico, literario, etc.—. Para poder ejecutar estas diligencias, será necesario que el programa cuente, por un lado, con un sistema de rasgos codificados referidos a la tipología textual y por otro, con una batería de rasgos lingüísticos que definan la adecuación de cada tipo de texto o estilo.

[12] Se sigue, en esta exposición, la clasificación y enfoque que se ofrece en Llisterri (2007) y Martí y Taulé (2011).

La descripción de estos tres tipos de programa junto con las técnicas que los posibilitan será desarrollada con más detalle en el siguiente capítulo.

Traducción automática

Unas de las primeras aplicaciones relacionadas con el ámbito de la LC que surgieron embrionariamente en los años 50 tenían por objeto la traducción automática.

En la actualidad son, posiblemente, las aplicaciones más populares que generan múltiples investigaciones y productos. No obstante, las inversiones de tiempo y recursos económicos no eximen a estos programas de presentar ciertas limitaciones en lo que se refiere a la gama de contenidos que pueden someterse a la traducción, y de afrontar ciertas dificultades que surgen en el trasvase de contenidos de lenguas tipológicamente distantes. En cualquier caso, existen aplicaciones profesionales que permiten obtener buenos resultados con textos especializados y sujetos a dominios bien delimitados. Se han desarrollado, además, sistemas de *traducción asistida por ordenador* que mejoran notablemente el trabajo del traductor.

Los problemas que surgen en este campo del PLN son los propios que involucra la comprensión de un enunciado. El sistema interpretativo que deba llevar a cabo la traducción, además de atesorar conocimientos lingüísticos —morfológicos, sintácticos, léxicos y semánticos— deberá contar con datos extralingüísticos que suelen agruparse bajo el título de *conocimiento del mundo*, información que difícilmente puede formalizarse. En los últimos años, para alcanzar los objetivos deseables en todo sistema de PLN, el ámbito de la traducción automática ha introducido técnicas estadísticas para evitar que todo el peso de la traducción recaiga en un complejo sistema de reglas. De esta inclusión surgen sistemas híbridos que combinan ambas metodologías, la codificación de reglas lingüísticas y las técnicas estocásticas.

La presencia en el mercado de estos programas es amplia y creciente. Algunos de los ejemplos para el español, desarrollados en España son Compendium, AutomaticTrans o InterNOSTRUM. Otros programas desarrollados en el marco europeo o América son Systran —incorporado en las herramientas de

Google y en el traductor Babel Fish, de Altavist—, Reverso, Websphere Translation Server o imTranslator, de la compañía Smart Link Corporation.

Recuperación y extracción de la información: minería de datos

El imparable aumento del reservorio de información alojado en la red, junto con la creciente digitalización de grandes fondos documentales, hace imposible el tratamiento manual de todos estos datos. Esta realidad ha generado la necesidad de disponer de un acceso automático a los datos contenidos en toda la información disponible. Surge, en este entorno, la minería de datos, entendida como el conjunto de técnicas y tecnologías que permiten explorar grandes cantidades de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto (Castro, 2012).

Estas técnicas ayudan a procesar el contenido de los datos, materia prima bruta, y facilitan que el usuario pueda atribuirle un significado o sentido para finalmente convertir esos datos en información. Con este fin, la minería de datos hace uso de prácticas estadísticas, algoritmos de búsqueda próximos a la Inteligencia Artificial y sistemas basados en el funcionamiento de las redes neuronales para lidiar con asuntos lingüísticos de diversa naturaleza.

Dentro de este contexto, se exploran técnicas para la recuperación y extracción de la información y se desarrollan motores de búsqueda, clasificadores de documentos por temas, sistemas que elaboran resúmenes automáticos de documentos o sistemas de análisis de sentimientos —suscitados por algo o alguien que suele patrocinar la investigación, *minería de opinión*—. Todas estas áreas, en la actualidad, se convierten en objetivos prioritarios en el ámbito de las tecnologías lingüísticas.

2.2 La Lingüística de Corpus

Las lenguas humanas, el objeto que se pretende procesar, no son directamente observables en su completitud y complejidad. Para ello el investigador tendría que estar no solo ante todas las emisiones que se han producido o que se producen,

sino también ante aquellas que pudieran ser producidas en el futuro, lo que nos lleva a una cantidad infinita e inabordable de enunciados.

Junto con el problema de la infinitud, convive el hermetismo de ciertos aspectos del lenguaje. Los principios que el investigador está tratando de descubrir y eventualmente codificar para que una máquina los interprete, el niño con dos o tres años ya los conoce intuitivamente y, para asombro del investigador, van más allá de la posibilidad de introspección consciente (Chomsky, 1988:14).

En este contexto, los corpus han sido a lo largo de la historia de la investigación lingüística, una herramienta vital para la formalización del lenguaje, un recurso esencial de la lingüística tradicional de corte empírico, de los estudios dialectológicos y de la lingüística histórica. No obstante, no ha sido hasta el desarrollo tecnológico reciente, cuando se han podido reunir, organizar y procesar mecánicamente los datos de la lengua con un volumen realmente significativo.

Así las cosas, como señala Rojo (2002) es ya casi un lugar común considerar que la lingüística basada en el análisis de corpus es la forma que adopta la lingüística descriptiva a finales del siglo XX y también que sus características básicas están fuertemente determinadas por la utilización de recursos informáticos.

En la aproximación tradicional hay una colecta selectiva de materiales llevada a cabo por expertos, mientras que en la basada en técnicas computacionales se manejan todos aquellos casos que cumplan con las condiciones especificadas en la expresión de consulta. Este resultado se le ofrece al investigador en bruto, sin ser filtrado ni seleccionado[13].

El corpus se convierte entonces, en el medio más imparcial para estudiar nuestro objeto de estudio entendido como lengua real, en uso, en contextos naturales —*actuación*— y no como abstracción teórica del sistema de una lengua —*competencia*—. El enfoque y metodología que ofrece el corpus es puramente empírico; los datos serán los que apoyen o contradigan una postura teórica, los que

[13] Como señala Rojo, los resultados de una consulta llevada a cabo con métodos de lingüística de corpus son en bruto, esto es, la ordenación o criterio que maneja la máquina para devolver los resultados es puramente mecánica y sujeta a los procedimientos de recuperación de la información; fecha del texto, orden alfabético de las palabras que estén situadas en el contexto inmediato, etc. (2002:108).

permitan inferir reglas y generalizaciones, los que proporcionen informaciones cuantitativas, etc.

Pero la recuperación de este método empírico, como sostienen Rafel y Soler, no habría podido producirse sin el desarrollo tecnológico que ha permitido la constitución y explotación de corpus cada vez más extensos y complejos (2003:43).

En un sentido amplio, estos vastos repertorios textuales se pueden definir como una colección de datos lingüísticos cuya selección y organización se lleva a cabo teniendo en cuenta criterios lingüísticos explícitos con el fin de ser utilizados como muestra de la lengua (Sinclair, 1996; Mairal Usón *et al*., 2012).

La consecuencia de todo ello, como puede anticiparse, es un cambio ostensible en la metodología, que provoca el despegue de una nueva disciplina; la Lingüística de Corpus, cuyo principal cometido es la descripción y análisis de la lengua con una base empírica, a partir de los datos que ofrecen los corpus dentro de entornos informáticos[14]. Parodi, en su introducción a la disciplina propone la siguiente definición;

> La Lingüística de Corpus constituye un conjunto o colección de principios metodológicos para estudiar cualquier dominio lingüístico, que se caracteriza por brindar sustento a la investigación de la *lengua en uso* a partir de corpus lingüísticos con sustrato en tecnología computacional y programas informáticos *ad hoc* (2008:95).

Para cualquiera de los enfoques metodológicos que se adopte en el área de la LC, el corpus será un punto de partida imprescindible para el desarrollo de aplicaciones basadas tanto en texto como en habla. Suelen ser la materia prima de lexicones y gramáticas computacionales, pilares necesarios para casi cualquier investigación. Los corpus son, en la actualidad, la fuente canónica de datos lingüísticos más importante con la que los sistemas de PLN pueden trabajar (Biber *et al.,* 1998; Periñán, 2012).

En un sentido "moderno" y más restringido, los corpus pueden ser definidos como:

[14] Para una presentación exhaustiva de la Lingüística de Corpus puede consultarse, además, Leech (1993, 2005); McEnery y Wilson (1996) y Biber *et al.* (1998).

> […] conjuntos de textos de lenguaje natural e irrestricto, almacenados en un formato electrónico homogéneo, y seleccionados y ordenados, de acuerdo con criterios explícitos, para ser utilizados como modelo de un estado o nivel de lengua determinado, en estudios o aplicaciones relacionados en mayor o menor medida con el análisis lingüístico (Santalla, 2005:45-46).

Sometidos a estas exigencias, los corpus informatizados han promovido avances significativos dentro de la Lingüística de Corpus, dando un nuevo impulso a los estudios descriptivos de diferentes aspectos de la lengua; prosodia, léxico, morfología, sintaxis, historia de la lengua, etc. Su automatización y parametrización ha sido igualmente provechosa en el campo del análisis gramatical automático de textos, tanto en sus aspectos morfológicos como en sus aspectos sintácticos.

Los planteamientos que propone la Lingüística de Corpus se basan, pues, en el *estudio de un todo,* el lenguaje y las lenguas, cuyo hábitat es infinito por su creatividad esencial, a partir del *análisis minucioso de una de sus partes*, que será siempre mínima en comparación con el objeto completo. Pero los corpus lingüísticos, sin cuestionar su valor, sufren la inevitable tara de la parcialidad y la incompletitud[15]. Como señala Rojo, un corpus no es un absoluto, y en consecuencia, debe ser concebido, tanto en su construcción como en su explotación, en tanto que muestra representativa de los parámetros, necesidades y fines de los que surge (2002).

2.2.1 Los requisitos

En este contexto, dentro del ámbito de lo computacional, los requisitos que se esperan de un corpus pueden sintetizarse en cinco premisas (McEnery y Wilson, 2001:21 y ss):

- Sus textos deben presentarse *en formato electrónico* para ser una herramienta útil al lingüista, de modo que sea posible automatizar tareas tales como la búsqueda y recuperación de información, los cómputos de frecuencias, la clasificación de los datos, etc.

[15] Para una definición, desarrollo histórico y principios de este enfoque se puede consultar Leech (1991), McEnery (2003), McEnery y Wilson (1996, 2001), Tognini-Bonelli (2001), Alcántara Plá (2007), Parodi (2008).

- Deben conservar la a*utenticidad de los datos*: los textos recogidos en el corpus deben ser muestras reales de uso de la lengua.
- Los textos recogidos deben obedecer a ciertos criterios *de selección* ya sean lingüísticos y/o extralingüísticos, para la finalidad concreta que persiga el corpus.
- Las *muestras deben ser representativas*: la selección de los textos debe responder, además, a parámetros estadísticos que garanticen que los textos "representan" la variedad de la lengua objeto de estudio. De este modo se minimiza el sesgo que todo corpus aloja inherentemente.
- Es necesario a*cotar su tamaño*: por lo general, los corpus constan de un tamaño finito, que se suele medir en millones de palabras —o formas— y que se fija antes de empezar la recogida de los textos.

El Corpus TIP, que será uno de los recursos sobre los que se apoya PatErr, cumple, como se verá en el Capítulo 4, con todos estos requisitos exigidos por las bases de la disciplina, lo que lo convierte en un corpus de referencia útil, fiable, representativo y potente.

2.2.2 Tipos de corpus

Hay muchos tipos de corpus, que pueden especializarse en función de su diseño, de las características formales o de los métodos utilizados para su constitución. En los últimos 30 años se han producido corpus de muy distintas lenguas y con muy diversas finalidades, y son sus fines, a la postre, los criterios que rigen los principios de diseño de cada corpus.

En primera instancia, puede establecerse una clasificación entre corpus escritos y corpus orales. Esta escisión refleja la diferencia entre los recursos y métodos desarrollados para las tecnologías del habla, y aquellos destinados al PLN. Junto con esta primera división, conviven múltiples taxonomías que se han diseñado para poner orden en el universo de los corpus posibles.

A grandes trazos, podemos proponer un panorama que diferencie cinco tipos básicos (Alcántara Plá, 2007)[16].

Los CORPUS ESPECIALIZADOS están integrados por textos que han sido seleccionados por presentar unas características concretas según el tipo de estudio que se quiera llevar a cabo. Normalmente la selección presenta particularidades que los alejan del uso general. Un ejemplo de esta tipología es el *Corpus Childes,* repositorio de transcripciones de las emisiones lingüísticas de niños de casi todas las nacionalidades, cuyas muestras incluyen valiosos datos sobre la adquisición del lenguaje, en casos de adquisición normal, en casos patológicos —afasias, dislalias, etc.— así como en casos de niños plurilingües.

Por otro lado, los CORPUS GENERALES O DE REFERENCIA representan inmensas compilaciones que deberían proporcionar una visión completa de la lengua. No están restringidos salvo en aspectos muy básicos como ser orales, escritos o mixtos. Los corpus de referencia deben recoger el máximo de datos para garantizar su *representabilidad, funcionalidad* y *variedad*. Representará, por lo tanto, todas las variedades de registro, tipos de discurso, de vocabulario, etc. de la lengua en cuestión. Entre sus aplicaciones primarias está el diseño y desarrollo de diccionarios y gramáticas. Algunas de las obras más relevantes son el *British National Corpus* [BNC] y el *Corpus de Referencia del Español Actual* [CREA][17].

Otro tipo de compilaciones son los denominados CORPUS COMPARABLES. Estas colecciones están compuestas por subcorpus que comparten todas las características básicas salvo la de la lengua o la variedad lingüística. Algunos de los más valiosos son el *International Corpus of English* [ICE] y el *C-ORAL-ROM* con representación, en el primer caso, de la diversidad lingüística del inglés, y en el segundo, de las lenguas romances en su versión oral.

Por su parte, los CORPUS PARALELOS están compuestos por corpus de textos idénticos en distintas lenguas. Pueden ser producidos simultáneamente o partir de un corpus original y elaborar el conjunto de subcorpus a partir de su traducción a

[16] Para una propuesta más amplia y pormenorizada puede consultarse la desarrollada en Torruella y Llisterri (1999) y Rafel y Soler (2003).

[17] Como se verá más adelante, el Corpus TIP se inscribe en este grupo ya que, en un sentido amplio, representan todas las variedades —diatópicas, diafásicas y diastráticas— posibles del español panhispánico escrito en la red.

otras lenguas. Se utilizan principalmente para el desarrollo de memorias de traducción o para aplicaciones relacionadas con la legislación internacional.

Por último, los CORPUS HISTÓRICOS contienen textos de distintos momentos históricos que se compilan para realizar estudios contrastivos de lingüística diacrónica. El *Corpus Diacrónico del Español* [CORDE] es un gran ejemplo de este tipo de trabajos.

Debemos mencionar que actualmente la red ofrece la posibilidad de acceder al conjunto más vasto de textos en formato electrónico que son muestras reales de uso de la lengua de todo tipo y materia y que pueden ser recuperados mediante cualquier buscador. Pero de acuerdo con los investigadores de estas disciplinas, *el valor de un corpus reside fundamentalmente en la calidad y fiabilidad de su codificación y sus anotaciones,* por lo tanto, la grandiosa hemeroteca de textos que internet nos ofrece, el llamado *big data,* necesita ser desenmarañado, letra por letra, organizado y catalogado, palabra por palabra, para poder convertirse en una herramienta útil tanto en la investigación lingüística —en todas sus posibilidades— como en PLN (Alcántara Plá, 2007:20).

2.2.3 Corpus para el PLN

Una última clasificación que puede establecerse, de máxima relevancia para las tareas que nos ocupan, se basa en la existencia de CODIFICACIÓN y ANOTACIÓN en el corpus. Según este rasgo de diseño los corpus serán simples, planos, no codificados ni anotados, o corpus codificados y anotados. Estos últimos están constituidos por textos a los que se les ha añadido manual o automáticamente etiquetas declarativas de algunos elementos estructurales de los documentos —título, principio de capítulo, pie de página, etc.—, codificación o etiquetas analíticas que recogen rasgos lingüísticos como la categoría gramatical, el modelo de flexión o el registro lingüístico —anotación—[18]. Sobre estos últimos se ha explicitado información lingüística presente ya antes de forma implícita. Este procedimiento,

[18] Teniendo en cuenta este aspecto, otro modo de clasificación frecuente es el que se inspira en el tipo de información que ha sido anotada. La tipología se corresponderá a los niveles lingüísticos estudiados tradicionalmente; *nivel acústico, fonológico, morfológico, sintáctico, semántico* y *pragmático*. Cada uno de estos niveles podrá formalizarse con diferente grado de precisión y matices, que estarán motivados por la base teórica y la finalidad sobre las que se sustente el trabajo.

la anotación, supone un proceso de caracterización lingüística de cada unidad registrada. La anotación, por lo tanto, añade valor a un corpus ya que lo habilita para responder a más preguntas sobre el funcionamiento del lenguaje[19].

En cualquiera de los dos procedimientos mencionados es esencial que las etiquetas empleadas en la codificación o anotación sean extratextuales, de modo que se puedan reconocer, modificar o eliminar con facilidad sin desestabilizar el resto del sistema. Para el desarrollo de estos procesos se recomienda, igualmente, acogerse a alguno de los estándares propuestos para estos proyectos, de modo que se garantice la portabilidad y reutilización de los textos incluidos en el corpus (Torruella y Llisterri, 1999). Debe mencionarse, a este respecto, el ambicioso y valioso trabajo del profesor Mark Davies que se concreta en el *Corpus del Español*. Este recurso recoge más de cien millones de palabras presentes en más de 20 000 textos del español de los siglos XIII al XX. El programa a partir del cual puede accederse al contenido, permite hacer búsquedas selectivas por palabras exactas o frases, categorías gramaticales, lemas, colocaciones, etc.[20]

A pesar de su utilidad, hay muy pocos corpus anotados de acceso libre para la comunidad científica, menos aún en nuestra lengua. La razón hay que buscarla, siguiendo a Civit, en el alto coste, tanto técnico como humano, que cobra la generación de un corpus enriquecido, especialmente si se pretende *que la anotación satisfaga ciertos criterios de calidad como la coherencia, la consistencia interna y la buena documentación, que son los que otorgan valor a la anotación* (2003;12-13).

Estos criterios de calidad se asentarán, en buena medida, en el establecimiento de una base gramatical sólida y coherente que permita justificar con consistencia y rigor cada uno de los datos que han sido anotados sobre el corpus.

En el Capítulo 5 se presentará un esbozo del marco teórico que recoge las cuestiones gramaticales esenciales para el desarrollo y fundamentación de la

[19] Algunas de las ventajas que se han señalado para los corpus anotados son: la facilidad de explotación, la reutilización y multifuncionalidad y la capacidad que ofrecen para llevar a cabo análisis explícitos (McEnery, 2003:454).

[20] Puede consultarse en http://www.corpusdelespanol.org/hist-gen/2008/x.asp

anotación de un lexicón que es finalmente la base sobre la que se llevan a cabo las tareas de anotación de los corpus.

2.3 La convergencia de disciplinas

Entendemos por método el conjunto de principios metodológicos generales del paradigma en el que se encuadre la investigación (Alcaraz Varó, 1990:7). Los paradigmas, en un sentido kuhniano, serán marcos de referencia, caracterizados por una homogeneidad relativa de pensamiento teórico básico que proporciona a la comunidad científica criterios para marcarse metas nuevas, seleccionar hechos relevantes que se conviertan en problemas de investigación y proponer soluciones a esos problemas dentro del paradigma (1990:11).

En este trabajo concreto debemos hacer ciertas precisiones. Inserto como está en un área multidisciplinar, cabe hablar de *dos métodos de aplicación paralela*; el que vertebra la parte teórica de *naturaleza lingüística*, propia del área de la gramática del español y la lingüística de corpus, y el que da sustento a la aplicación práctica, que es de *naturaleza computacional*. Ambos métodos tendrán un horizonte común, un producto lingüístico concreto, el error, que será el objeto de examen empírico, investigación teórica y formalización en este trabajo. De esta prioridad surge como marco concreto lo que hemos dado en llamar una LINGÜÍSTICA DE ERRORES CON FINES COMPUTACIONALES.

Por otro lado, pensar en el lenguaje *para* la computación obliga a tener presente una realidad: los métodos y conclusiones adoptados en la investigación lingüística tienen como fin la implementación en un sistema computacional, con las concesiones y límites que este implica. Es necesario, pues, tener siempre presente el grado de viabilidad —posibilidad de implementación a través del ordenador— que ofrecen las conclusiones formuladas.

Para presentar a grandes trazos los métodos de análisis lingüístico que se observarán y se desarrollarán en los epígrafes sucesivos, podemos acudir a la tradicional diferencia entre el método inductivo y el hipotético deductivo, que según Givón refleja en nuestros estudios la dicotomía que ha existido entre las dos

tendencias opuestas que han dominado la epistemología occidental; el empirismo y el racionalismo (1989:20).

De acuerdo con Popper, no creemos que ningún método pueda atribuirse la exclusividad científica de la investigación, por lo que se adoptará una actitud ecléctica, esto es, *inductivo-deductiva.* Parret, llama a este método híbrido *par procedimental* que está formado por los dos métodos y se comporta inductiva o deductivamente según las conveniencias o necesidades de la investigación, evitando los corsés que impone la adhesión inquebrantable a una u otra declaración de principios (1974).

Se utilizará el *método empírico* siempre que el trabajo adopte como herramientas los corpus o masas de datos lingüísticos para la descripción lingüística. El método se asentará en la observación de los comportamientos de las unidades dentro del sistema, que descubrirán los atributos de la realidad lingüística del español actual.

Por otro lado, y como es práctica común en el *paradigma deductivo*, se tomarán ciertos axiomas por indubitados y se pondrán en tela de juicio las conjeturas y formulaciones derivadas de la observación de los hechos lingüísticos, que buscarán confirmación, refutación o reajuste de reglas.

La diversidad de visiones y aproximaciones al lenguaje no implica su incompatibilidad, muy al contrario, encontramos, para este tipo de investigación de PLN, complementariedad entre diversos métodos de evaluación lingüística, de modo que las soluciones a los problemas que planteará la lengua en un entorno computacional, vendrán dadas por métodos asociados a diferentes paradigmas de investigación.

Las fuentes bibliográficas a las que se ha recurrido para el análisis lingüístico son, como se verá, gramáticas y tratados sobre el español desarrollados en el seno de diversos paradigmas. En los epígrafes dedicados a la anotación del Lexicón TIP, Capítulo 5, y al desarrollo de PatErr, a partir del Capítulo 6, se expondrán las líneas esenciales que vertebran la justificación lingüística de este trabajo.

La fuente de referencia en lo que a la norma del español se refiere será la *Nueva gramática de la lengua española* —en adelante NGLE—, obra que se ha

convertido en la primera gramática académica desde 1931 y en la gramática panhispánica por excelencia. Toma como punto de partida la tradición, a la que incorpora las nuevas investigaciones de la lingüística moderna. Como todas las gramáticas, la NGLE es una obra analítica que trata de reflejar la forma de expresarse de los hispanohablantes. Su mérito, como señala Cervera Rodríguez, está en la pretensión de aunar enfoques y aspectos diferenciados evitando contradicciones, como el de la unidad dentro de la variedad de usos, para lograr una obra cohesionada, integrada y coherente (2011:14). La NGLE será en esta investigación una herramienta esencial para el cotejo, sistematización y formalización de los datos.

Pero la lingüística computacional y computacionable debe hacer uso de métodos legibles para un ordenador, máquina que presenta una gran capacidad para registrar, gestionar y manipular datos a gran velocidad, pero que carece de capacidad interpretativa del objeto de estudio en cuestión. Su falta de intuición lingüística reclama la elaboración de sistemas simbólicos coherentes, formalizados y legibles que, asentados sólidamente sobre bases de datos y/o algoritmos, ofrecerán a la máquina materia sobre la que obrar.

En este sentido, los métodos de ingeniería lingüística propios de la tarea computacional que se utilizarán para sostener el aparato teórico tendrán que ver, por un lado, con el uso de las expresiones regulares en programación que derivará en el desarrollo de un lenguaje de codificación propio[21] y, por el otro, con la estructura informática que recogerá toda la información contenida en PatErr que se asienta sobre una base de datos. A partir de estos dos componentes se llevará a cabo una tarea de formalización y codificación de los resultados de la teorización lingüística acerca de los errores del español escrito.

2.4 ¿Dónde se inscribe esta propuesta?

A partir del panorama general que hemos presentado, podemos adscribir el presente trabajo que se actualiza en el diseño y desarrollo de un recurso para la

[21] Friedl (2006).

revisión textual del español —PatErr—, en el área concreta del PLN, que asume los proyectos y productos derivados de las tecnologías relacionadas con el texto y que a su vez participa de la disciplina más genérica de la LC.

Siendo este el entorno que propicia el enfoque, las herramientas y los objetivos de esta propuesta, la metodología y planteamientos deberán buscarse en la lingüística de corpus, y el objeto de investigación y tratamiento, el error, será un producto de la lengua española que deberá observarse desde los alrededores del sistema y con los condicionamientos propios que impone la máquina.

Debe tenerse en consideración además que, como sucede en otros dominios de estudio, el PLN tiene un componente teórico, de ciencia básica, la LC en su rama teórica y una vertiente más aplicada y tecnológica, que hemos descrito bajo el título de Ingeniería Lingüística, encargada de desarrollar sistemas concretos, *ingenios,* relacionados con el uso de la lengua.

Uno de los trabajos que se presenta en esta tesis, el diseño de un etiquetario para un lexicón de español, tiene que ver con ese componente teórico derivado de la gramática española. El otro trabajo, que se concreta en el diseño y desarrollo de un sistema basado en un depósito de errores codificados para la revisión automática de textos, debe inscribirse en la otra cara de la moneda, en la que se ocupa de soluciones concretas para problemas reales en los que la materia que debe procesarse es la lengua fuera de las abstracciones del sistema.

Capítulo 3

Estado de la cuestión

Uno de los lances recurrentes en el campo de las tecnologías lingüísticas es el enfrentamiento entre los requisitos de la teoría lingüística y los propios de los sistemas computacionales que exigen trabajar eficazmente con márgenes exiguos de tiempo. El PLN es, al fin, una disciplina de corte aplicada orientada a la consecución de productos materiales que puedan integrarse en el mercado tecnológico. A este respecto, la ingeniería informática denuncia con frecuencia la debilidad de las herramientas desarrolladas para procesos automáticos y el escaso interés de buena parte de las teorías lingüísticas por producir gramáticas computacionables, esto es, que puedan ser procesadas. Los lingüistas, por su parte, ponen en entredicho el recurso a las soluciones *ad hoc* sin fundamentación lingüística que suelen ser habituales en las labores prácticas de ingeniería. En cualquier caso, parece evidente que el tratamiento automatizado de las lenguas debe surgir de un cuidadoso equilibrio entre teoría lingüística y formalización.

Las aplicaciones de tecnología lingüística se fundamentan en una infraestructura de recursos tecnológicos de realización costosa y compleja. Dichos recursos tienen una base tanto informática como lingüística: los programas, los lenguajes de representación y el diseño de las aplicaciones, por un lado y los datos lingüísticos representados en forma de léxicos, gramáticas, bases de conocimiento, etc., por el otro.

Los programas informáticos suelen ser independientes de la lengua y, por lo tanto, concentran una mayor masa crítica de investigadores que, además, pueden compartir los conocimientos de manera mucho más general. En cambio, los especialistas que se dedican a la elaboración de estructuras de datos o recursos lingüísticos —bases léxicas, gramáticas, corpus anotados—, que son necesariamente dependientes de la lengua, se encuentran más atomizados y, sin embargo, el esfuerzo y el coste de desarrollo que requieren puede ser idéntico o superior al de los programas informáticos (Martí y Taulé, 2011).

Ambos perfiles profesionales deben afrontar una tarea compleja como es el procesamiento del lenguaje, que se caracteriza por su infinitud —el número de frases posibles de una lengua es abierto— y su complejidad: es sintomático que no exista todavía una teoría universalmente aceptada sobre el lenguaje humano.

Los problemas principales con que deben enfrentarse las aplicaciones lingüísticas, tanto en la modalidad oral como escrita, son fundamentalmente:

- la ambigüedad intrínseca al lenguaje que emerge desde los elementos fónicos hasta el significado;
- la variación lingüística, es decir, la posibilidad de expresar una misma idea de maneras distintas
- y la creatividad del lenguaje humano, esto es, la creación constante de nuevos términos y expresiones que hace de las lenguas entidades dinámicas en un proceso dialéctico de cambio constante.

Si se tiene en cuenta que las aplicaciones de tecnología lingüística tratan la lengua viva, el uso real del lenguaje en contextos de comunicación muy variados, se pone de manifiesto la dificultad que entraña el desarrollo de tales aplicaciones.

3.1 Aproximaciones y modelos en el PLN

Es común, como se ha intentado poner de manifiesto, la separación (dramática) entre el conocimiento puramente lingüístico y los algoritmos que lo procesan. Históricamente los dos enfoques adoptados en la investigación para la resolución y desarrollo de estas aplicaciones que abordan en el procesamiento y tratamiento de las lenguas han sido el SIMBÓLICO y el ESTADÍSTICO. Junto a estos, conviven, en la actualidad, el ENFOQUE HÍBRIDO, que adopta los métodos de los paradigmas clásicos y el modelo CONEXIONISTA, que se basa en criterios derivados de la información disponible sobre la actividad neurológica en el procesamiento del lenguaje (Liddy, 2001).

3.1.1 Modelo simbólico

En una primera aproximación, el enfoque simbólico puede caracterizarse por estar basado en el conocimiento lingüístico que se presenta recogido en forma de reglas abstractas de carácter introspectivo. Para su explotación se crean algoritmos que operan con estas estructuras de datos simbólicos cuyo contenido representa las descripciones gramaticales de una lengua.

Este enfoque está basado en las aproximaciones chomskianas al paradigma lógico formal, cuyo objeto era especificar, de manera rigurosa y explícita, la estructura de una lengua mediante formalismos gramaticales. Como cabe sospechar, estas gramáticas se basan solo en la competencia del hablante, no en la actuación, por lo que las reglas gramaticales formalizadas solo se aplicarán si se satisfacen todas las condiciones que derivan de la competencia.

Su falta de realismo le impide, en muchas ocasiones, abordar las lenguas en movimiento, esto es, en el uso, en el que surgen insospechados giros lingüísticos, configuraciones focalizadas modelizadas por la intención y el contexto comunicativo y, en última instancia, el error.

3.1.1.1 *Las gramáticas computacionales*

Una gramática es una descripción premeditadamente finita de un lenguaje potencialmente infinito. Este proyecto, por lo tanto, solo se consigue con la aplicación de reglas y elementos recursivos. Una gramática computacional, por su parte, se concibe como esa descripción de conocimiento lingüístico formalizada pero que, además, pueda ser empleada tanto como una herramienta de análisis automático como en el funcionamiento de algunas de las aplicaciones de PLN.

A pesar de la diversidad de aproximaciones a la implementación de gramáticas computacionales dentro del paradigma simbólico, muchos de los esfuerzos se han centrado en encontrar el formalismo más adecuado para representar la información lingüística que, además, sea adaptable a cualquier lengua. Estos formalismos gramaticales o metalenguajes son, pues, lenguajes que sirven para la descripción de otros lenguajes y, aunque comparten objetivos, difieren en el modo de conceptualizar una misma descripción sintáctica del

funcionamiento de la lengua y del sistema de interacción entre sus diferentes módulos —léxico, morfológico, sintáctico, etc.—. Estas diferencias de conceptualización se deben, en gran medida, a que cada formalismo está condicionado por una teoría gramatical diferente. Algunos de los lenguajes de programación más invocados para el desarrollo de gramáticas computacionales son PATR o ALEP (Gómez Guinovart, 2001)[22] .

En cuanto a las teorías y modelos que inspiran estas gramáticas, las opciones que han concitado la mayor parte de los proyectos e investigaciones en los últimos años son las conocidas como GRAMÁTICAS DE UNIFICACIÓN y las GRAMÁTICAS DE RESTRICCIONES (Moreno Sandoval, 2001; Llisterri, 2007).

Las Gramáticas de Unificación nacieron en los años 80 como solución a los problemas de procesamiento en tiempo real de los modelos gramaticales previos. Estos paradigmas simbólicos están fuertemente influidos por las teorías lexicistas del paradigma generativo de la época[23] y reducían el papel de la gramática en aras de dotar de mayor importancia al léxico. La mayoría de estos modelos lexicistas se fundamentan en el supuesto de que la estructura argumental de un verbo está directamente determinada por sus propiedades léxicas, o dicho de otro modo, la entrada léxica de un verbo determina su comportamiento sintáctico[24]. La filosofía de estas gramáticas es, en síntesis, codificar la máxima información posible en el nivel léxico[25].

Los modelos más explorados e implementados que incorporan estos planteamientos son la *Gramática de Estructura Sintagmática Generalizada*, GPSG, la *Gramática Léxico Funcional*, LFG, y la *Gramática de Estructura Sintagmática Nuclear,* HPSG (Moreno Sandoval, 2001; Periñán, 2012; Ramírez González, 2013).

[22] PATR (*Parsing and Translation*), es un mecanismo desarrollado a mediados de los 80 en la Universidad de Standford. Su finalidad es escribir gramáticas sintagmáticas formadas por reglas independientes de contexto y aumentadas con estructuras de rasgos sobre las que opera la unificación. Una gramática descrita en PATR consta de un conjunto de reglas y un lexicón. (Para una exposición más detallada de estos formalismos, puede consultarse Moreno Sandoval, 1998 y 2001).

[23] Este paradigma estaba ya desprovisto de reglas transformacionales y eludía la separación de la estructura profunda y la superficial, en favor de un fortalecimiento del componente léxico y el (Periñán, 2012:23).

[24] Esta tendencia de identificar los diversos argumentos que configuran la estructura argumental del verbo a partir de la semántica del evento, que arrancó con el *aktionsart* de Vendler (1967), es la que adopta la mayor parte de las teorías de la lingüística actual, (Van Vallin y La Polla, 1997; Levin y Rappaport, 2005; Van Vallin, 2005). *Cf.* en Mairal Usón y Gonzálvez García (2010).

[25] Un caso práctico de implementación de estos formalismos gramaticales lo encontramos en el corrector para español y griego moderno *GramCheck* (Ramírez Bustamante y Sánchez León, 1996).

Las Gramáticas de Restricciones, *Constraint Grammars,* abanderadas por F. Karlsson (1995) parten de la anotación de todas las posibles funciones sintácticas que pueda desarrollar una palabra para realizar después un proceso de desambiguación y seleccionar la función más adecuada en una oración concreta. Son aún, sistemas demasiado precarios y no ofrecen la solvencia de los modelos de unificación. No obstante, son buenas herramientas para tareas de desambiguación y análisis superficiales (Balari, 1999; Rodríguez, 2000).

En cualquiera de las opciones expuestas, para el desarrollo de estas gramáticas debe considerarse que una de las características de los sistemas de verificación basados en ellas es, como se anunció, que deben analizar secuencias/oraciones tanto correctas como incorrectas. Por este motivo, las gramáticas que actúen de base para los programas de verificación o corrección de textos deben contener tanto la gramática de la competencia, como la gramática de la actuación; deben ser *robustas*[26], es decir, capaces de procesar lo no gramatical o lo no registrado, sin que sus sistemas colapsen al enfrentarse a secuencias insospechadas.

3.1.2 Modelo estadístico

El enfoque estadístico, por su parte, palía la rigidez de las reglas lingüísticas partiendo de amplias muestras de lenguaje real. Estos datos lingüísticos suelen estar alojados en los corpus, que son etiquetados y utilizados para crear modelos y patrones estadísticos que capturen el funcionamiento de las lenguas. El objetivo de estos sistemas es la inferencia de conocimiento a partir de los datos reales que se obtienen de la actuación de los hablantes en busca de regularidades significativas. Una vez recolectados los datos deben catalogarse y anotarse para proceder al cálculo de frecuencias de cada unidad.

Un caso intuitivo de resolución de problemas sobre bases estadísticas puede observarse en los procesos de desambiguación o de corrección ortográfica

[26] La *robustez* es una propiedad deseable en todo sistema informático o *software*. Los sistemas robustos se caracterizan por ser capaces de mantener sus condiciones esenciales de desarrollo pese a recibir perturbaciones o ruidos. Es, pues, la capacidad de un sistema de absorber anomalías o variaciones impredecibles y seguir en funcionamiento.

automática. La palabra *ademas,* —segunda persona del singular, del presente de indicativo de verbo *ademar*— convive en el lexicón junto al adverbio *además*. Cualquier corrector ortográfico comercial corrige automáticamente la forma verbal sin tilde y la sustituye por el adverbio. Esta decisión automática no está motivada por una batería de reglas sintagmáticas restrictivas que aportan información sobre la categoría gramatical susceptible de contener un error —verbo-adverbio—, sino por un estudio de frecuencias sobre un corpus que revela la escasa presencia del verbo *ademar* en contraste con la frecuencia del adverbio. Asentada en esta lógica, la máquina, nutrida por los datos estadísticos, asume que detrás de *ademas* hay un error ortográfico y no una forma flexionada del verbo *ademar*.

3.1.2.1 *N-gramas*

En la actualidad, recogiendo la tendencia de los últimos años, es muy frecuente desarrollar estos modelos basados en la estadística mediante la técnica de N-gramas. Se entiende por N-grama *una subsecuencia de n elementos consecutivos de una secuencia dada.* En el caso que nos ocupa, los elementos son las palabras y la secuencia, el texto que se pretende tratar. Este tipo de planteamientos parten del contexto local que rodea a cada palabra para extraer frecuencias, colocaciones y concordancias[27].

Para llevar a cabo esta estrategia, un programa agrupa linealmente las palabras de una secuencia de dos en dos, de tres en tres, de cuatro en cuatro, etc. Así, de la oración *La ausencia de lluvias trajo graves consecuencias*, podemos extraer los siguientes trigramas:

la ausencia de
ausencia de lluvias
de lluvias trajo
lluvias trajo graves
trajo graves consecuencias

[27] Adoptamos la definición de McEnery *et al*. sobre la *concordancia*: [...] *is a list of all of the occurrences of a particular search term in a corpus, presented within the context in which they occur –usually a few words to the left and right of the search term* [...] *As with collocations, concordances provide information about the 'company that a word keeps* (2006:42).

A partir de estos trigramas se comprueba si cada una de esas combinaciones aparece en un corpus de amplia cobertura de textos escritos; si se registra en el corpus menos veces de lo esperable —según los análisis estadísticos previamente ejecutados—, la secuencia es susceptible de contener un error o un desvío.

El problema que presenta esta técnica es que limita la verificación solo a los errores que se presenten en un contexto de contigüidad. Si el error excede los contextos más inmediatos de un elemento pasará desapercibido para el programa, que se ejecutará exclusivamente a nivel local.

Los programas de revisión textual que cuentan con esta técnica como recurso pueden tomarla bien como complemento a los programas que proyectan un etiquetado y análisis morfosintáctico sobre el texto para llevar a cabo su función correctora —programas de metodología híbrida—, bien como planteamiento fundamental que a partir de información estadística inspire la lógica y funcionamiento del programa.

3.1.3 Modelo conexionista

El modelo conexionista se basa en la utilización de modelos de redes neuronales para representar el conocimiento lingüístico de un modo más realista. El objetivo de este método es simular la actividad neuronal que suponemos que se desarrolla en el cerebro para el procesamiento de la lengua.

En un modelo conexionista no están explícitos ni una gramática, ni un lexicón. Su *modus operandi* se basa en el reconocimiento de estructuras de diferentes niveles —fonemas, morfemas, oraciones, etc.—, que se realiza sobre la base de semejanzas en los patrones de activación de nodos. Así dos estructuras serán similares si se excitan los mismos nodos. Esta aproximación, que trata de simular la acción de nuestros recursos cognitivos, se acerca en muchos puntos al enfoque estadístico; en primer lugar, hace uso de estas técnicas —ponderaciones y asignación de pesos según frecuencias— y, en segundo, trata con muestras reales del lenguaje, con el producto de la actuación, no con las abstracciones regidas por la competencia, como es el caso de los modelos simbólicos.

Las TÉCNICAS HÍBRIDAS, por su parte, combinan uno o más de los modelos anteriores, con el fin de complementar las ventajas de cada uno y resolver problemas de dominios y aplicaciones específicos.

Las necesidades que fueron modelando estos planteamientos, aunque empiezaron a vislumbrarse en los años 40, no reclaman una solución contundente hasta los años 80. Es entonces cuando se perfilan, establecen y especializan estos paradigmas. Ofrecemos, a continuación, una breve reseña de carácter histórico que ayuda a componer el mapa de necesidades, soluciones y bamboleos que se observan a lo largo de la historia reciente del PLN (Periñán, 2012:24 y ss.).

3.2 Cronología de las aproximaciones del PLN

En la década de los 80 se vivió dentro del paradigma simbólico una corriente lexicista que derivó, dentro del PLN, por un lado, en las Gramáticas de Unificación y por otro lado, en el auge de los modelos probabilísticos, empleados principalmente en las tecnologías del habla, el etiquetado gramatical, el análisis sintáctico y la interpretación semántica. Mientras que el enfoque simbólico era utilizado para tratar los problemas más significativos del PLN, los planteamientos estadísticos solo servían de complemento al enfoque simbólico, y aún no estaban consolidados como un paradigma autónomo para el PLN.

En los años 90 se constató un resurgir de las teorías empiristas, alentadas en cierto modo, por el desarrollo de la Lingüística de Corpus y los avances en *hardware*, memoria y velocidad de procesamiento. A estas circunstancias, cabe añadir otra definitoria; la irrupción de internet, que facilita la accesibilidad no solo del conocimiento sino también de los recursos lingüísticos.

Este empirismo iba más allá del análisis de los datos lingüísticos, y se centró en la aplicación de métodos estadísticos al PLN. Como señala Periñán, *el paradigma estadístico fue convirtiéndose progresivamente en el estándar de numerosos campos del PLN* (2012:24) debido, entre otras cosas, a las ya mencionadas debilidades que presentaba el paradigma simbólico. Este se mostraba incapaz de proporcionar de forma flexible un tratamiento adecuado a un *input* defectuoso o una realización

lingüística nueva. Los sistemas basados en estadística, en cambio, se demostraban más robustos para estos casos, siempre que hubieran sido entrenados con volúmenes de datos anotados suficientemente significativos.

El racionalismo de los métodos basados en la codificación manual de reglas dio paso a métodos probabilísticos y de aprendizaje automático sobre corpus de los que estos sistemas pueden adquirir automáticamente el conocimiento lingüístico.

Como cabe imaginar, los años 90 presenciaron un cambio de paradigma en el sentido que propone Kunh, del simbolismo a la estadística, de los sofisticados prototipos de laboratorio basados en complejos modelos formales que, a la postre, no satisfacían las necesidades del mundo real, a los trabajos de ingeniería lingüística y las soluciones inmediatas y realistas. Este auge no obsta para que se abandone totalmente el enfoque simbólico, pero provoca un cambio en las prioridades investigadoras y comerciales.

Desde los 90 hasta la fecha, el enfoque estadístico ha dominado las investigaciones de PLN. Esta elección ha sido propiciada en gran medida por la accesibilidad a los recursos computacionales que permiten un tratamiento de las lenguas en un contexto real partiendo de vastos corpus de datos reales sobre los que se trabaja en búsqueda de la *mayor efectividad posible incluso a expensas de una clara fundamentación lingüística teórica* (Moure y Llisterri, 1996).

De hecho, como señala Periñán (2012), la mayoría de las ponencias en congresos de PLN tratan sobre soluciones de ingeniería a problemas prácticos; las investigaciones actuales en ingeniería lingüística no se fundamentan en la lingüística, sino en las estadísticas y la teoría de las probabilidades que dan soluciones más simples, aunque menos plausibles o elegantes desde un punto de vista teórico. Un ejemplo de aplicación basada en métodos probabilísticos son los traductores asociados a *Google*, que no se fundamentan en ninguna teoría lingüística, ni en componentes gramaticales sino en los datos que surgen a partir del análisis estadístico de los textos.

En cualquier caso, como señala este investigador, los enfoques simbólico y estadístico pueden coexistir perfectamente con el fin de desarrollar un sistema

más robusto, y suplir las carencias de un modelo con las ventajas del otro (Periñán, 28:2012).

Para completar esta panorámica sobre el estado de estas cuestiones y para ceñir el ámbito en el que se inscribe este trabajo, es necesario acotar el lugar de los correctores en las industrias de la lengua, sus técnicas, sus aproximaciones y sus niveles de intervención.

3.3 Sistemas de verificación textual; los correctores

Es frecuente, en la escasa bibliografía que se ocupa del desarrollo de estos sistemas, distinguir tres niveles susceptibles de corrección de complejidad creciente: la verificación ortográfica, la verificación gramatical y la verificación de estilo. Si la corrección ortográfica se centra en la palabra, entendida como secuencia de caracteres entre espacios en blanco, en la corrección gramatical se tienen en cuenta las combinaciones de más de una palabra, normalmente, fragmentos de la oración no siempre contiguos, mientras que en la corrección estilística es la oración el elemento central de procesamiento.

Las herramientas desarrolladas más utilizadas y extendidas son los correctores ortográficos. Por su parte, los verificadores gramaticales y de estilo —muchos aún en fase de desarrollo— tienen un menor nivel de aceptación y eficacia y levantan aún ciertas suspicacias en el usuario.

3.3.1 Dos aproximaciones para abordar el problema

A partir de las publicaciones consultadas, cabe vislumbrar un par de enfoques metodológicos que hacen uso de diferentes recursos para afrontar la revisión y corrección documental dependiendo del nivel objeto de la verificación.

Obviamente, ambos tipos de procedimiento o tipos de análisis parten de un etiquetado y anotación de las palabras o segmentos involucrados en el error. Estos dos grandes bloques de recursos pueden sintetizarse como sigue. (Ramírez Bustamante *et al*., 1997; 1998).

3.3.1.1 *Técnicas de bajo nivel*

Los recursos que llevan a cabo estas técnicas actúan en el nivel de la palabra, del signo. Su acción se encamina a la segmentación de unidades, el análisis morfológico y, en algunos casos, a la desambiguación de elementos, esto es, la adjunción de la etiqueta categorial propia dentro de un contexto.

Las técnicas de bajo nivel funcionan en el paradigma, y su ámbito de aplicación suele ser la corrección ortográfica y tipográfica, procedimientos que se llevan a cabo en el paso previo a la verificación gramatical, en un estadio de preprocesamiento, entendido este como un análisis previo a la interpretación sintáctica y semántica.

Las herramientas básicas que permiten este tipo de análisis son los LEMATIZADORES, los FLEXIONADORES, los ETIQUETADORES y algún sistema DESAMBIGUADOR. Como puede intuirse, la potencialidad de este tipo de análisis depende en gran medida de la lengua a la que se aplique. Señalan a este respecto Ramírez Bustamante y Sánchez León que 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 la contextual (1997:155).

A pesar de las evidentes limitaciones de estas técnicas, su acción no tiene por qué limitarse al nivel ortográfico; estos procedimientos típicos del preprocesamiento pueden servir, además, como apoyo a la corrección gramatical en casos y contextos locales y de contigüidad en los que se hace innecesario un análisis más profundo porque la irregularidad detectada se puede solventar con técnicas de bajo nivel de abstracción o métodos estadísticos como los N-gramas.

En casos como los que siguen, puede advertirse esta posibilidad de sortear análisis complejos para llevar a cabo una corrección de nivel gramatical:

**Me niego ha hacer reformas.*

**Se estableció un pacto de acuerdo a sus relaciones comerciales.*

**Habían fuertes presiones.*

Estas secuencias erróneas comparten el hecho de que el elemento incorrecto, a nivel gramatical, puede capturarse verificando su contexto-cotexto[28] más inmediato mediante, por ejemplo, el análisis de N-gramas. Un estudio de estos revelará que la estructura «*ha* + infinitivo» es ajena, por su escasa frecuencia, al repertorio configuracional de nuestra lengua y, por lo tanto, posiblemente sea portadora de un error. Lo mismo sucederá con **de acuerdo a*.

A partir del cotexto que rodea a la forma *ha* o al segmento **de acuerdo*, que algunos autores llaman *condiciones de error* (Ramírez Bustamante *et al.,* 1997) se pueden extraer patrones de error más abstractos que pueden funcionar como patrones fijos para la corrección. Así ante una secuencia abstracta como «*ha* + infinitivo», el programa puede automatizar la corrección «*a* + infinitivo».

En el último ejemplo, se observa un caso que ilustra la actual corriente de hacer concordar el verbo *haber* con valor existencial con el sustantivo del complemento directo, despojando a este verbo de su naturaleza impersonal; **Habían fuertes presiones*. A partir de una revisión superficial sobre un corpus del funcionamiento real de estas formas en el discurso, se puede concluir que cualquiera de las variantes flexionadas en plural del verbo *haber* seguidas de cualquier elemento de la lengua —determinantes, sustantivos, pronombres, adverbios, preposiciones, conjunciones, etc.— a excepción de un participio, deberán ser corregidas a su versión impersonal, es decir, en singular.

Con el uso de estas técnicas de bajo nivel se consiguen resultados satisfactorios en términos de teoría lingüística y de eficiencia computacional, prescindiendo de análisis de máxima abstracción que pasan por el uso, y previo desarrollo, de algún tipo de gramática computacional robusta capaz de procesar tanto lo correcto como lo erróneo.

Un proyecto que ha obtenido resultados favorables a partir de la utilización de estos recursos es el desarrollado por Ramírez Bustamante *et al.* (1998); *CON-TEXT*. Se trata de un corrector gramatical basado en técnicas de bajo nivel que parte de un procesamiento a nivel morfológico y se complementa con la aplicación

[28] El sentido que se le atribuye a este concepto a lo largo de estas páginas es el de *conjunto de elementos lingüísticos que preceden o siguen a un elemento lingüístico en el texto*, ya este una palabra, un sintagma o toda una frase. Quedan excluidos, pues, otros aspectos típicamente relacionados con el contexto de la situación comunicativa que se apartan de lo estrictamente textual.

de un repertorio de reglas lingüísticamente motivadas que *proporcionan puntos de anclaje fiables y suficientes para la construcción de un prototipo de verificación gramatical* (Ramírez Bustamante *et al.*, 1998:166).

Su predecesor, *GramCheck* (1996), basado en técnicas de alto nivel, reveló la necesidad de revisar algunas cuestiones de diseño y método con el fin de superar ciertas limitaciones e imposiciones onerosas asociadas a estas técnicas de alto nivel, en favor de otras técnicas más ligeras e igualmente fiables. *CON-TEXT* se basa en una simplificación de los métodos computacionales usados tradicionalmente en la verificación gramatical basada en el conocimiento lingüístico. Aunque esta práctica suponga sacrificar la cobertura del programa final, *CON-TEXT* se muestra eficiente en casos en los que las incidencias se presentan en contextos locales.

Debe inscribirse en el entorno de la utilización de estas técnicas de bajo nivel la propuesta que aquí se ofrece basada en el recurso PatErr, que aprovecha la potencialidad de estos sistemas de análisis lingüístico básico para la codificación de patrones de error, su posterior interpretación y explotación por parte de autómatas y la ulterior identificación de estos en el texto que se coteja.

3.3.1.2 *Técnicas de alto nivel*

Los recursos de alto nivel operan en el eje sintagmático para llevar a cabo la verificación gramatical. Estos métodos trabajan con agrupaciones de elementos, por lo que el programa requiere un mayor nivel de abstracción en el análisis, en este caso, de orden sintáctico. El término alto nivel, por lo tanto, se aplica no solo por la constatable complejidad de los mecanismos computacionales para tratar los 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* (Ramírez Bustamante y Sánchez León, 1997:148).

Es necesario pues, disponer de una gramática a partir de la cual se detecten las posibles violaciones cometidas y una vez identificadas poder ofrecer una opción satisfactoria de acuerdo con esas reglas legítimas previamente codificadas. Intuitivamente puede deducirse que la resolución de problemas con este tipo de técnicas resulta mucho más costosa tanto en términos de trabajo humano, como en términos de tiempo de procesamiento y ejecución.

3.3.2 Y un problema más; la ambigüedad

Debe hacerse una mención en este punto a un problema que constriñe, limita, traba y define el estado de la cuestión del PLN; la ambigüedad, que está presente y amenazante en el desarrollo de cualquier proyecto que tenga como materia el procesamiento y tratamiento de las lenguas.

La ambigüedad surge en el uso cuando es posible más de una interpretación para una determinada palabra, secuencia u oración. Tradicionalmente se establecen tres tipos de ambigüedad o niveles en los que puede concurrir;

Ambigüedad léxica: se encuentra en el eje paradigmático, cuando una palabra permite más de una interpretación. Se asocia con fenómenos como la polisemia —*corriente*, *cubo*—, la homonimia —*vale*, verbo y sustantivo— o la ambigüedad categorial —*este, aquella,* adjetivos determinativos y pronombre demostrativos—.

Ambigüedad estructural: surge en el sintagma, cuando una oración permite más de un análisis sintáctico, lo que implica tener más de una lectura para una misma secuencia; *Habló a sus estudiantes de informática.*

Ambigüedad semántica: ocurre cuando es posible proyectar más de una interpretación semántica sobre un elemento en la frase, en el sintagma. En *Pedro está en su despacho y Lola también* cabe una doble interpretación a partir del *también,* que puede tener un valor identificativo —mismo espacio— o distributivo —cada uno en su despacho—.

El primer paso para una desambiguación global inequívoca de las lenguas es, obviamente, el relacionado con las categorías gramaticales. Hasta que cada signo lingüístico no se adhiere a una única categoría gramatical, su significado tanto en el paradigma como en el sintagma y el texto es un elenco de posibilidades flotantes. Los ejemplos citados son casos de ambigüedad categorial que deben ser resueltos por la máquina; *vale, a, de, estudiantes, informática, despacho,* son términos que pueden desarrollar más de una categoría gramatical. Para cerrar el significado global de cada frase será necesario, en primer lugar, etiquetar cada palabra con el membrete categorial que en ese caso concreto está ejerciendo, pero será necesario que previamente —en el proceso de anotación del lexicón— se tenga previsto todo el repertorio de categorías que cada forma puede adoptar. El análisis gramatical en

estos estudios ha de ser, por lo tanto, exhaustivo y minucioso allí donde el hablante no lo es porque solo contempla, en cada caso, los análisis que tienen sentido en el contexto lingüístico y comunicativo en que se produce la interacción.

El objetivo de un desambiguador será pues, el de asignar a cada palabra la categoría más *apropiada*, dentro de un contexto. Es decir, dada una secuencia de palabras, dotada cada una con el conjunto de etiquetas posibles que pueda desarrollar, el desambiguador deberá devolver una secuencia de etiquetas que sea la más *verosímil* dado un contexto. Ante las dificultades que se plantean, es común el desarrollo de desambiguadores que, aunque no resuelven totalmente el problema, son capaces de eliminar las opciones imposibles o menos probables y ofrecer las más plausibles. No obstante, habrá casos en los que solo la intervención del humano que atesora conocimientos semánticos, pragmáticos e intencionales sobre la situación comunicativa podrá elucidar el análisis correcto.

Independientemente de las técnicas, recursos y consideraciones que deben observarse para el desarrollo de programas de corrección, lo cierto es que las páginas web, las aplicaciones de los teléfonos móviles, el correo electrónico, los buscadores y prácticamente todos los programas que requieren un uso de la lengua, incorporan técnicas de revisión textual automática. A continuación, se revisarán los tipos y alcance de los sistemas que se han desarrollado en el área de la verificación y corrección automática de textos, con independencia de cuál haya sido su metodología, planteamiento y destino final.

3.4 Correctores ortográficos

Son múltiples los proyectos que a partir de 1960 han tratado el tema de la corrección ortográfica. Estos estudios y tentativas han propiciado el surgimiento y fortalecimiento de aplicaciones que permiten la corrección ortográfica, ya sea como tarea final o como parte del flujo de trabajo de algún otro sistema —traducción automática, minería de datos, extracción de información, etc—).

Los correctores ortográficos suelen estar alojados en los procesadores de textos e intentan identificar los errores ortográficos en el documento para, a continuación, sugerir la posible corrección o directamente automatizarla.

La técnica informática más habitual para detectarlos consiste en la comparación de las palabras presentes en el documento con una lista de palabras correctas almacenadas en un lexicón, donde además de estar registradas las formas canónicas, están presentes todas sus posibilidades flexivas y derivativas. De este modo, tras el análisis contrastivo, el corrector ortográfico señalará la presencia de un error cuando una palabra del texto no encuentre réplica en el lexicón de partida.

Los errores de ortografía que se cometen durante la escritura a través del ordenador pueden derivarse de dos hechos; el desconocimiento de la norma lingüística o el despiste-lapsus del escritor. A los errores habituales que comete un hablante nativo frente al papel, deben sumarse nuevos errores que surgen por el uso de la computadora, bien sean de manos del escritor, bien por el uso de herramientas para la digitalización como un escáner o el procedimiento de digitalización OCR. Las faltas generadas por el desconocimiento de la norma en algún punto del proceso de escritura se consideran técnicamente como *errores de competencia*, mientras que los segundos, que devienen de descuidos y pueden ser considerados erratas, caen en el saco de los *errores de actuación*.

A pesar de que el repertorio de errores de competencia fluctúa ampliamente de un usuario a otro, existen determinados factores lingüísticos que favorecen su aparición; la falta de correspondencia entre una grafía y la fonética de una palabra —*zigurat, ahínco*, etc.—, las posibilidades de segmentación que ofrecen algunas formas —*porque, por qué*—, las vacilaciones que emergen de los pares homófonos —*echo, hecho*— o las discrepancias que surgen entre la norma y el uso —⊗*sólo*, ⊗*guión* o ⊗*truhán*—.

Por otro lado, los errores de actuación pueden derivar de errores mecanográficos o simplemente de una distracción o falta de atención. A propósito de estos, y como premisa inicial, debe tenerse en cuanta que, según numerosos estudios empíricos, en la escritura asistida por ordenador son muy escasos los

casos en los que el error se ubica en la primera letra de una palabra (Gómez Guinovart, 2000:238).

Tras la identificación de una secuencia errónea, el programa pasa al proceso de corrección. La técnica clásica aplicada en la mayoría de estos casos consiste en partir de la palabra errónea y comprobar qué mecanismo de error se ha fraguado y cómo puede solucionarse mediante los cuatro mecanismos formales que estadísticamente recogen toda la casuística:

- inserción de una letra; **amrillo*
- elisión de una letra: **amarilllo*, **hamarillo*
- sustitución de una letra por otra: **amarillp*, **amariyo*
- trasposición de letras adyacente: **amairllo*

Se aplican estas técnicas sobre la palabra errónea y se cotejan sus resultados en el lexicón buscando formas correctas entre las palabras que comiencen por la misma primera letra y que solo suponga un tipo de mecanismo —una sola operación— inserción, elisión, sustitución o trasposición de una letra. Si la búsqueda resulta infructuosa, el corrector ampliará su ámbito de búsqueda a palabras que comiencen por una letra diferente —**smarillo* por *amarrillo*— o a aquellas que se puedan derivar de más de un tipo de error —**smrillo* por *amarillo*—, con elisión añadida en este caso.

Aunque los resultados de los correctores ortográficos son muy aceptables, su grado de precisión no es absoluto. No son pocos los casos en los que el programa corrector indica la presencia de un error donde no lo hay porque la palabra buscada, a pesar de ser correcta, no está recogida en el lexicón. Esto es muy común con los nombres propios, los neologismos, los tecnicismos, los términos propios de un sublenguaje y una gran cantidad de palabras que se cuelan a diario en nuestro repertorio disponible. Esta carencia, no obstante, se puede solventar fácilmente ampliando el diccionario de partida empleado en el procesador.

En otras ocasiones, la solución no es tan inmediata debido a que el error cometido da lugar a una palabra legítima, ortográficamente correcta, diferente de la pretendida; recuérdese el caso ilustrativo del verbo *ademar*. Sucede, igualmente, con expresiones que involucran a las series de homófonos o casi homófonos

—*vaya/ valla/ baya*, *aya*/*haya/halla*, etc.—, o con pares que admiten varias opciones de segmentación como —*sino/si no*—. En estos casos, si no se dispone de un analizador sintáctico, lo común es que la incorrección pase desapercibida para la máquina debido a que la palabra portadora del error tiene presencia en el lexicón de partida. Estos ejemplos son solo algunos de los muchos posibles que ilustran los focos de error ortográfico que puede presentar un texto.

3.5 Correctores gramaticales

En comparación con los correctores ortográficos, el ámbito de aplicación de un revisor gramatical es mucho más impreciso y, por lo tanto, su fiabilidad se compromete. Aunque a diario se evidencian importantes líneas de mejora, este tipo de correctores constituye una buena herramienta de ayuda a la escritura. Algunos de los errores que comúnmente detectan y solventan están relacionados con problemas de concordancia gramatical, con usos incorrectos de la puntuación, con la detección de oraciones demasiado largas o con la presencia abundante de preposiciones o conjunciones relativas.

El grado de abstracción requerido para estos análisis hace que sea necesario dotar al sistema de conocimiento lingüístico más complejo, lo que significa, por un lado, invertir en recursos humanos para la formalización de ese conocimiento, y por otro, asumir una ralentización en el tiempo de ejecución del programa. A diferencia de la corrección ortográfica, que puede detectarse y corregirse con relativa facilidad, la solvencia de un analizador sintáctico es mucho menor. No es una tarea fácil para la máquina —ni para el lingüista computacional— decidir de manera automática que una secuencia desarrollada por palabras ortográficamente legítimas —presentes en un lexicón— contienen un error sintáctico ya que, entre otras razones, como señala Gómez Guinovart,

> [...] las indicaciones recogidas en las gramáticas normativas nunca son tan exhaustivas como para albergar todos los tipos de enunciados que debe manejar un procesador de textos (2000:239).

Aunque la utilización de componentes de tecnología lingüística como analizadores morfológicos y sintácticos o gramáticas computacionales pueden mejorar la calidad del resultado como ocurre en el caso de la corrección ortográfica, los sistemas de revisión gramatical también provocar falsas alarmas de error. Para paliar estas carencias se han pensado métodos alternativos capaces de afrontar estas tareas. Actualmente la técnica informática más utilizada para la identificación de los errores gramaticales surge del enfoque, asumido por esta propuesta, de reconocimiento de patrones de error.

Los patrones de error son secuencias de al menos una palabra que recogen pautas, elementos y/o contextos que contienen un error o desvío con respecto a las descripciones gramaticales que ofrece la norma. El reconocimiento se lleva a cabo a partir de una búsqueda que hace el programa, que recorre todo el texto corregible, de esos patrones desarrollados que definen la cobertura de la aplicación. Aunque esta pueda parecer una operación trivial, el establecimiento de unos patrones de validez general para un determinado tipo de error requiere, en ocasiones, un cierto grado de abstracción lingüística complementado con el uso de información morfológica.

Esta técnica puede ser utilizada como recurso o metodología principal o como complemento para sistemas que tratan desviaciones gramaticales y están basados en el análisis sintáctico de alto nivel, programas estos que ralentizan la ejecución y el procesamiento debido al coste computacional en términos de comparación de rasgos y estructuras lingüísticas (Ramírez Bustamante *et al.*, 1997:148).

Para esta segunda opción, aproximación complementaria que se subordina a otras técnicas de alto nivel o gramáticas computacionales, puede considerarse la sugerencia que Ramírez Bustamante y Sánchez León proponen de incluir, en su caso en las bases léxicas de *GramCheck* y *CON-TEXT*, las formas incorrectas de aquellas variantes ortográficas, sean flexivas o no, de casos relacionados con la flexión que implican que un proceso morfológico regular se aplica a un lema que pertenece a un paradigma flexivo irregular. Es el caso, por ejemplo, de formas tan

extendidas como **andé*[29], que es el resultado de una regularización en la flexión de formas irregulares (1997).

Como es lógico, los resultados de esta técnica dependerán de la amplitud y precisión de los patrones establecidos para nutrir al programa. Este únicamente detectará un error cuando se corresponda con alguno de los patrones previstos, pero no todos los errores gramaticales son fácilmente previsibles ni acotables en un contexto susceptible de *patronizar*; es por esto que la verificación gramatical por patrones precisa complementarse con otras técnicas auxiliares, bien de alto nivel, como los análisis sintácticos automáticos para solventar casos de discordancia intesintagmática, bien con procedimientos estadísticos como los que parten de la técnica de N-gramas.

3.6 Correctores de estilo

Otra herramienta de ayuda la escritura desarrollada en el ámbito del procesamiento de textos es el corrector estilístico. En general, este tipo de verificadores realiza la función de comprobar la adecuación de un texto a un conjunto de reglas previamente definidas y seleccionadas por el usuario, que deberá anticiparle al programa algún parámetro como la tipología textual, el registro, el estilo o la variedad lingüística.

La misión de estos programas, generalmente basados en estudios empíricos y en el uso de técnicas estadísticas, es constatar que los rasgos lingüísticos del documento analizado sean afines con los trazos lingüísticos que el sistema tenga preestablecidos como preceptivos para la categoría textual y parámetros seleccionados. Estas categorías textuales están constituidas por un conjunto de rasgos lingüísticos formales que verifican la presencia o ausencia en el texto de determinados giros o expresiones.

[29] Ningún corrector que trabaje con coeficientes de similitud entre formas propondría la forma correcta *anduve*, que es muy diferente a la incorrecta**andé*. Las propuestas basadas la similitud serían *ande, anda, ando, nadé, nade*, etc. todas ellas más cercanas a la forma incorrecta, pero fracasaría en la captura de la falta de conocimiento que se esconde detrás de este error (Ramírez Bustamante *et al.,* 1997).

Otra de las técnicas más comunes de verificación estilística es la evaluación del nivel de legibilidad del texto, esto es, el grado de complejidad del texto determinado por ciertos factores lingüísticos cuantificables, como la extensión de las oraciones, la longitud de las palabras o la cantidad de preposiciones de una frase. Para llegar a alcanzar buenos resultados a partir de este enfoque se requiere disponer, cuando menos, de una tipología textual y una enumeración lo más detallada posible de los rasgos lingüísticos que caracterizan a cada tipo de texto o estilo, —literario, técnico, administrativo, etc.—.

Aunque el punto de desarrollo de estos sistemas no permite el diseño de programas globales y autónomos, sí hay aplicaciones parciales y de ámbito restringido que resultan de gran utilidad para los escritores y traductores. Por otro lado, la mayoría de los correctores disponibles para el español incorporan algún tipo de tratamiento estilístico dentro de su repertorio de medidas correctivas.

3.7 Algunos correctores desarrollados para el español

Muchas de las aplicaciones inscritas en las que hemos denominado industrias de la lengua se desarrollan o culminan en un entorno comercial, por lo que la parte teórica sobre la que se basan estas herramientas no es publicada ni está disponible para el investigador. Por otro lado, las publicaciones que se encuentran sobre proyectos llevados a cabo en el seno de los grupos de investigación universitarios son bien aproximativos, revelan cuestiones de perspectiva y arquitectura básica sin profundizar en el desarrollo y soluciones concretas previstas para la aplicación, bien conclusivos, exponiendo estudios contrastivos con otros productos o empíricos en términos de cobertura. No obstante, de estas publicaciones se desprenden conocimientos válidos que han ayudado al planteamiento general de esta propuesta.

Algunos de los primeros proyectos de corrección ortográfica y gramatical para el español que se llevaron a cabo con la participación activa de lingüistas fueron los desarrollados por Ramírez Bustamante y Sánchez León en el seno de la Universidad Autónoma de Madrid. Esta participación se concretó en los proyectos

ya citados *GramCheck* (1996), demostrador de sistema de corrección gramatical para hablantes nativos de español y griego moderno y *CON-TEXT*, corrector gramatical de bajo nivel (1998).

Por las descripciones que ofrecen los autores en los artículos en los que publican los resultados de estos proyectos, —no se ha tenido acceso al programa o recurso en sí—, se conoce que *GramChek* está basado en procedimientos de corrección gramatical de alto nivel que requieren una descripción también de alto nivel de las estructuras gramaticales. La ejecución del programa se apoya en una gramática altamente especificada desarrollada con formalismos de unificación basada en rasgos con tipos. Los errores detectados y corregidos por este demostrador son los relativos a la concordancia en género y número —inter e intrasintagmática—, y la adición, sustitución y omisión de preposiciones regidas en complementos oblicuos y directos —preposición *a*—. También realiza correcciones a nivel paradigmático de errores relacionados con la flexión. En los casos en los que se registra un error o deficiencia, el programa ofrece mensajes, sugerencias y una corrección automática según la tipología de error detectada.

El proyecto *CON-TEXT* (1998) se basó en la anotación morfológica para llevar a cabo la tarea de revisión textual. Los autores lo definen como *corrector gramatical basado en el uso de técnicas de bajo nivel*, tales como la segmentación, el análisis morfológico y la ulterior desambiguación.

La elección de estas técnicas básicas en el preprocesamiento de un texto está motivada porque estas son menos onerosas que aquellas relacionadas con análisis de alto nivel —análisis sintácticos— y sirven para detectar una gran variedad de errores que pueden ser tratados a bajo nivel sin necesidad de llevar a cabo un análisis completo de las oraciones. Con el uso de estos métodos alternativos, se logra, por un lado, la simplificación de los métodos computacionales utilizados tradicionalmente, y por otro, la creación de herramientas complementarias útiles, capaces de ejecutar una verificación gramatical y de estilo. *CON-TEXT* se compone de un conjunto de reglas locales que describen secuencias erróneas de descripciones morfosintácticas enunciadas declarativamente que se contrastan con las descripciones morfosintácticas del texto anotado.

Un programa de corrección automática que está plenamente inserto en las industrias de la lengua y ofrece unos resultados satisfactorios es *Stilus*, de la empresa Daedalus, que para su módulo gramatical aprovecha las reglas y algoritmo desarrollados por el proyecto *CON-TEXT* (Villena *et al.,* 2002).

Otro sistema de corrección pensado para los aprendientes de español como lengua extranjera se concretó en *CorrectMe*, desarrollado por un consorcio de investigadores de la UNED, cuyos resultados se presentan en San Mateo (2006). A diferencia de los reseñados proyectos, este trabajo se basa en la técnica de N-gramas. San Mateo propone un corpus de bigramas utilizado como corrector ortográfico y gramatical que sirva como herramienta destinada a mejorar las habilidades de escritura durante el aprendizaje del español. En lugar de utilizar el sistema de etiquetado y análisis sintáctico que utilizan gran parte de estos programas, su propuesta se basa en partir de análisis estadísticos que se basan en la comparación de combinaciones de palabras con un corpus de referencia de 100 millones de palabras. Bajo este planteamiento, se señalan los pares de palabras poco o muy poco frecuentes, y que, generalmente lo son porque contienen un error.

Partiendo de este mismo enfoque estadístico y del mismo tipo de usuarios, deben mencionarse los trabajos de Nazar y Renau en los que se utilizan estas técnicas estadísticas apoyadas en grandes corpus. En Nazar y Renau (2012) hacen un estudio sobre el español utilizando como corpus el contenido de *Google Books* para construir una base de datos, en principio correctos, con pentagramas que tienen presencia como mínimo de 40 veces dentro del corpus. A partir de estos datos, es posible desarrollar una herramienta de ayuda para los estudiantes no nativos de español.

Asociado a estos recursos automáticos para el aprendizaje del español, cabe mencionarse el proyecto *ELE*, *Tutor Inteligente*, (Ferreira y Kotz, 2010). Este programa o STI, *sistema tutorial inteligente*, se asienta sobre procedimientos y prácticas propios de la Inteligencia Artificial y utiliza técnicas de comprensión y generación de lenguaje natural. Su finalidad, en palabras de sus autoras, es *facilitar los procesos de aprendizaje personalizados.* Para ello, se requiere una recopilación

previa de los errores que recurrentemente aparecen en los textos de aprendices de una lengua extranjera.

Por otro lado, debe hacerse mención a un proyecto desarrollado para el aprendizaje de la lengua inglesa, el corrector gramatical *E-gramm.* Diseñado por el grupo de Tecnologías de la información y sistemas de la UNED (2008), hace uso de la técnica de N-gramas y las aplica sobre un corpus de redacciones en inglés de las que se extrajeron los errores que típicamente cometen los estudiantes en el proceso de aprendizaje de esta lengua. En una fase posterior al desarrollo del prototipo se incluyeron recursos apoyados en el uso de expresiones regulares que demostraron ser de gran utilidad para la detección y corrección de errores dentro del texto corregible en lengua inglesa (Chacón Beltrán, 2008).

Finalmente, debe señalarse que en la red pueden encontrarse algunas demostraciones de correctores para el español como el de *SIGNUM* (proveedora de *Microsoft*), *DataSpell* de *Bitext* o el que compone el pack de herramientas COES de la Universidad Politécnica de Madrid. En cuanto a la corrección ortográfica, parcela que más desarrollos ha tenido, son conocidos los correctores ortográficos de la familia de *Spell*, desarrollados en el entorno Linux (*Aspell*, *Ispell*, *Hunspell*) y el asociado al proyecto de Nadasdi y Sinclair, *BonPatron* (2001) cuya versión española es *SpanishChecker*.

3.7.1 Un sistema basado en la identificación de patrones

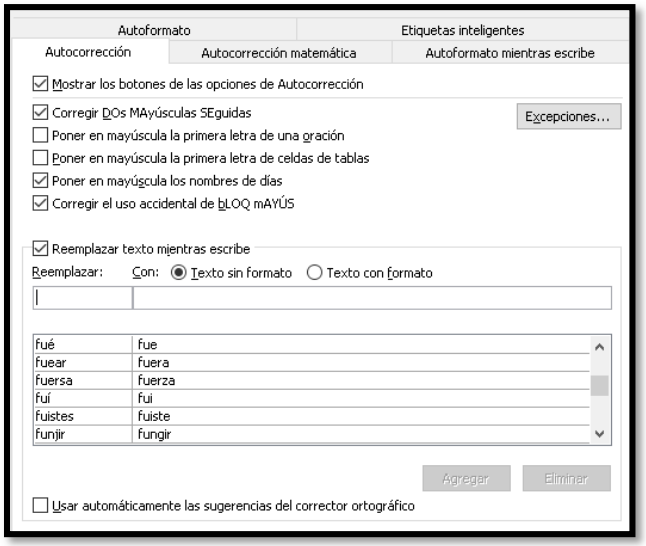
Por último, se reseñarán algunos aspectos de un modelo de corrector comercial, que parte del enfoque que se adoptará para el desarrollo de PatErr, la identificación de patrones de error.

El corrector ortográfico y gramatical del español alojado en el programa de edición de *Microsoft Word* es uno de los más difundidos y utilizados. El sistema parece contener dos módulos que presentan diferente grado de eficacia, precisión y cobertura. Mientras el corrector de ortografía se considera una herramienta muy útil para la revisión ortográfica, capaz de afrontar tanto los errores de competencia como los de actuación, el corrector gramatical no se ofrece aún como una herramienta fiable y provoca suspicacias en el usuario.

Como se señaló en la Introducción, el enfoque de corrección a partir del reconocimiento de patrones de error previamente compilados en una base de datos parece materializarse, al menos en parte, en el sistema de este corrector automático. Como también se puso de manifiesto, la ausencia de publicaciones que aborden y expresen los planteamientos y metodología concretos que se adoptan para el desarrollo de estos patrones obliga a aplicar técnicas deductivas sobre el producto final para obtener información que permita conjeturar sobre los entresijos y criterios en los que se asienta este programa.

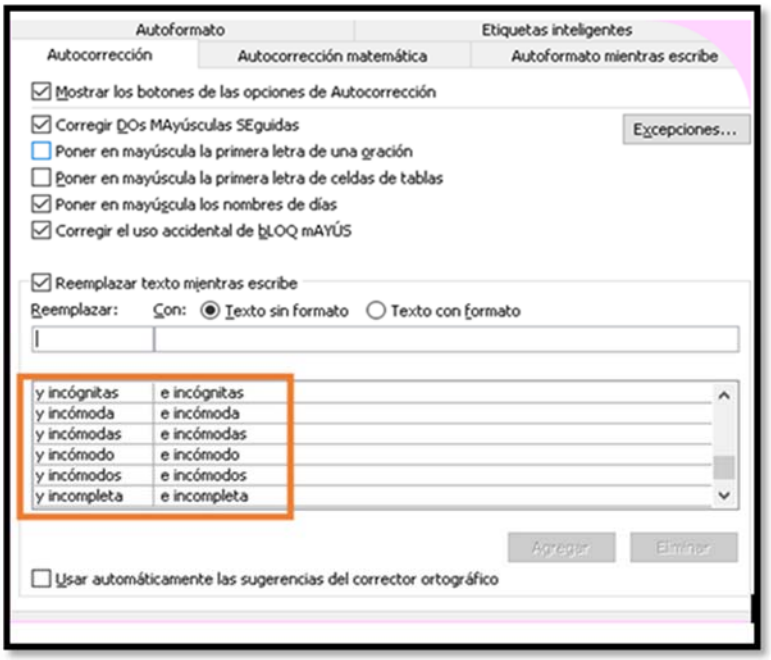
Tras las exploraciones llevadas a cabo sobre la arquitectura accesible de este corrector, se han observado dos estrategias paralelas para el desarrollo del proceso de corrección; por un lado, el corrector contiene un módulo gramatical y otro de estilo que solventan algunos casos básicos de concordancia, dequeísmo, usos incorrectos de modos verbales, usos de mayúsculas, laísmo, etc. Este hecho invita a pensar que tras el módulo gramatical hay algún análisis sintáctico automático de alto nivel de abstracción.

Por otro lado, se observa que en las *opciones de autocorrección* —corrección automática— que presenta contienen una batería de unos 500 patrones de error. La naturaleza del error contenido en estos patrones es variada, como puede observarse en la siguiente figura:



Estos patrones desarrollan contenidos tanto ortográficos, como gramaticales, de segmentación, ortotipográficos, etc.

A pesar de contar con cerca de 500 patrones, el número de casos que podrá ser resuelto se reduce considerablemente. La identificación y tratamiento automático de un mismo error, en ocasiones, requiere el desarrollo de varios patrones diferentes para copar la cantidad de flexiones que puedan ofrecer las formas involucradas en el error.



A partir de esta exploración y de la comprobación de las inconsistencias ya expuestas que presenta este programa en el proceso de corrección automática lo que parece evidente es que para superar estas carencias los patrones de error deben, idealmente, captar toda la casuística de desvíos o incorrecciones que puedan surgir en el uso de la lengua.

La innovación que supone el planteamiento que inspira el desarrollo del depósito de patrones PatErr es, como se concretará a lo largo de estas páginas, el diseño de patrones de error codificados capaces de encapsular en un solo patrón, siempre que la lengua lo permita, toda la casuística que pueda devenir de un mismo fenómeno o estructura errónea.

Estos patrones, procesados por técnicas de bajo nivel como las expuestas en Ramírez Bustamante *et al.* (1998), permitirán desarrollar un sistema de revisión

textual para el español capaz de abastecer a un programa final que ofrezca las garantías y cobertura exigibles en un entorno comercial.

Capítulo 4

Recursos lingüísticos para el PLN

> Las tecnologías lingüísticas pretenden facilitar el uso de las computadoras y el acceso a las redes que configuran la sociedad de la información y del conocimiento, sin necesidad de renunciar a nuestro uso habitual y natural del lenguaje.
> Para el desarrollo de estas tecnologías es necesario contar con una batería de recursos lingüísticos que se convierten en elementos esenciales para el desarrollo de cualquier aplicación, ya esté relacionada con el análisis o con la generación de textos. Estos recursos suelen agruparse en tres categorías; bases de datos léxicas, corpus y herramientas de análisis lingüístico[30].

Para el planteamiento, diseño y desarrollo del recurso PatErr han sido necesarios algunos componentes propios del ámbito del PLN; un lexicón anotado, un corpus de frases y un conjunto de herramientas de análisis lingüístico de bajo nivel; un lematizador, un conjugador y un flexionador.

El lexicón es una de las bases fundamentales de cualquier recurso enfocado a la ingeniería de aplicaciones relacionadas con el PLN y la interacción entre la máquina y el hombre. De su extensión y fiabilidad depende buena parte del éxito en el procesamiento del lenguaje que lleven a cabo los programas a los que dé soporte. La mayoría de estos, indefectiblemente, se basan y hacen un uso extenso y extensivo del lexicón, que ha de estar correctamente diseñado, estructurado e implementado además de permitir un acceso fácil y rápido a la información que contiene (Moreno Ortiz, 2000).

El corpus de frases ha sido, por su parte, la fuente de información básica para el trabajo propio de la lingüística de errores. Como se ha manifestado, un corpus permite, de modo empírico, extraer información sobre el funcionamiento real del lenguaje a partir de muestras de su uso. A partir de este recurso se han podido

[30] Llisterri (2007).

llevar a cabo exploraciones y generalizaciones, corroborar hipótesis y, finalmente, elaborar patrones de uso y de error.

Las herramientas de análisis, por último, han sido necesarias para establecer el diseño de los patrones del repositorio de errores y el del lenguaje que los codifica. Para la explotación del recurso que presentamos, los patrones deberán ser interpretados mediante automatismos por estas herramientas propias del preprocesamiento y el tratamiento morfológico.

4.1 Las bases de datos léxicas

La materia prima con la que trabajan los programas dedicados al PLN son las palabras. Estas se compilan en amplios lexicones codificados y anotados que dan lugar a lo que se conoce como LEXICÓN COMPUTACIONAL que, a diferencia de los diccionarios convencionales, contiene la información morfológica, sintáctica y semántica[31] relevante para copar tres funciones dentro del campo de acción que se aborda en este trabajo;

- desarrollo de aplicaciones o sistemas dedicados a las tecnologías del texto,
- incorporación del lexicón a las herramientas de análisis automático,
- anotación de corpus textuales.

La importancia y centralidad del lexicón en las aplicaciones de PLN es un hecho indubitado. Como afirma Calzolari,

> It is almost a tautology to affirm that a good computational lexicon is an essential component of any linguistic application within the so-called 'language industry', ranging from NLP systems to lexicographic enterprises (1994:267).

Un lexicón computacional, es pues, un repositorio de información léxica desarrollado con el objeto de servir de soporte representacional a diversas

[31] La representación del significado de cada unidad léxica suele almacenarse en las denominadas *ontologías*. Siguiendo a Periñán, [...] *mientras el lexicón contiene el conocimiento lingüístico, la ontología estructura jerárquicamente el conocimiento del mundo compartido por un hablante medio* (1:2005). El diseño de la ontología corresponde a una fase posterior del tratamiento, para aplicaciones que precisen la interpretación semántica del aducto. El lexicón, en cambio, es el punto de partida de cualquier aplicación.

aplicaciones de PLN. Una de sus características deberá ser, por lo tanto, la multifuncionalidad, la capacidad de ser reutilizado para distintos fines sin que sea necesario modificar su estructura o contenido. Para conseguir esto, será crucial alcanzar el mayor grado posible de independencia entre los datos, que deberán estar codificados y anotados con un formato computacionable. Esta información se revelará a los sistemas que operen con él mediante una etiqueta unívoca que se asocia a cada lema.

El PROCESO DE ANOTACIÓN, de darle contenido a las etiquetas, tiene una serie de requerimientos básicos que podemos sintetizar en tres puntos fundamentales (Leech, 1993, 2005; McEnery y Wilson, 1996):

I. Decidir el *tipo de información* lingüística que se va a anotar.
II. Asumir una *perspectiva teórica* que fundamente todo el proceso de anotación.
III. *Especificar un modelo o guía de anotación* que dé cuenta de cómo se ha formalizado la lengua y cuáles han sido los principales problemas lingüísticos de la anotación.

Con respecto al primer requisito, el tipo de información que se pretende registrar, se pueden identificar al menos cinco tipos de conocimiento que son relevantes en cualquier sistema de PLN (Moreno Ortiz, 2000: 2.2).

- *Conocimiento fonológico*, acerca del sistema de sonidos y la estructura de las palabras, los patrones de acentuación, la entonación, etc.
- *Conocimiento morfológico*, sobre la estructura de las palabras. Se codifica información como la relativa a las marcas *-s* y *-es*, que en español se añaden a los sustantivos para formar el plural.
- *Conocimiento sintáctico,* sobre las configuraciones de las palabras en el eje sintagmático. Se incluyen rasgos como la intransitividad verbal, que ofrece información sintáctica sobre los constituyentes de la oración.
- *Conocimiento semántico,* acerca del significado y sentido de las palabras. Estos rasgos, además, podrán revelar datos sobre la sintaxis; un concepto como el que representa *pintar* implica la existencia de

dos entidades, un agente —el que pinta— y un paciente —la cosa que el agente pinta—. De esta información podrá derivarse la estructura argumental de esta forma.

- *Conocimiento pragmático,* claves relacionadas con el contexto que posibilitan la interpretación de un texto. Se anotan datos que facilitan el análisis de las presuposiciones del hablante, de las intenciones comunicativas o de las implicaturas que subyacen en una frase o contexto en el que aparece la palabra etiquetada.

Aunque toda la información contenida en los lexicones puede ser relevante y esencial según el tipo de procesamiento que prevea el programa para lo utilice, será la adscripción a una clase gramatical la que hará que emerjan el resto de las propiedades; que *casa* sea una u otra categoría —sustantivo o verbo— dará lugar a que surjan unas u otras propiedades en nuestra conciencia lingüística. Según la categoría que desarrolle la forma presentará un haz de rasgos diferente.

Independientemente del tipo de conocimientos o dosis de información que se codifiquen, estas deben hacerse explícitas en el lexicón y servirán de puntos de anclaje que ayuden a los programas a seleccionar las piezas y procesarlas según los diferentes niveles-necesidades de tratamiento. Por ejemplo, una gramática computacional ineludiblemente se formulará a partir de categorías gramaticales; adherir esta información a cada elemento facilitará a los programas ejecutores el reconocimiento y manipulación de las piezas léxicas.

En cuanto a la CONSTRUCCIÓN DE UN LEXICÓN COMPUTACIONAL, puede llevarse a cabo a través de dos métodos; la *creación* y la *adquisición* (Calzolari y Picchi, 1994). El primer método implica la construcción manual del repertorio a partir de la introspección del lexicógrafo. El segundo permite su elaboración de forma automática o interactiva aprovechando y reutilizando los recursos lingüísticos que se contienen en formato electrónico. Uno de estos recursos básicos, sin duda el más utilizado para la compilación léxica, es el diccionario electrónico que ofrece una gran cantidad de información lingüística contenida en las entradas léxicas.

La metodología basada en la competencia del lingüista parece insuficiente para la elaboración de un lexicón; el inventario activo y disponible de un hablante, por más competente que sea, no representa la diversidad léxica de las lenguas. Esta

metodología presenta, como se advierte, problemas tanto teóricos como prácticos. Por una parte, una misma realidad lingüística puede ser interpretada bajo distintos enfoques. Por otra, la consistencia del lexicón se tambalea cuando su tamaño excede de unos cientos de entradas. A toda esta falta de sistematicidad y consistencia, debe señalarse el gran coste humano en horas de trabajo que implica la creación de un lexicón computacional a gran escala.

Para evitar este inconveniente, es común hacer uso de ambas estrategias; la automática en primera instancia y la manual, que auxilia a la automática y solventa sus posibles deficiencias.

4.1.1 Teoría de la anotación

Como se viene sugiriendo, la anotación lingüística es una forma de enriquecimiento de un corpus textual que se realiza mediante la introducción de etiquetas o membretes descriptivos o analíticos que explicitan información lingüística implícitamente contenida en dicho corpus (McEnery, 2003). Las posibilidades de explotación de la información que manejan los recursos lingüísticos para el PLN —corpus, lexicones, ontologías, etc.— dependen pues, en gran medida, del etiquetado y anotación de los que se les ha dotado.

La tarea de anotación es una práctica interpretativa, producto de la comprensión humana de un texto que no siempre se corresponderá con la interpretación igual de válida llevada a cabo por otro analista. No hay una fórmula objetiva ni mecánica que decida qué etiquetas deben aplicarse a determinado fenómeno lingüístico o qué fenómenos lingüísticos deberán ser codificados en un corpus, lo que implica, en algunos casos, optar por una interpretación de las varias posibles.

Para hacer expresos los fundamentos sobre los que se basa la anotación será necesario establecer un ESQUEMA DE ANOTACIÓN, con independencia del nivel que se pretenda abordar, que deberá constar al menos de dos apartados imbricados (Leech, 1993, 2005);

- Un conjunto de etiquetas morfosintácticas que se adjuntarán a cada unidad de análisis del lexicón o corpus. Este conjunto será el etiquetario o *tagset*.
- Un conjunto de criterios que permitan determinar la categoría de cada unidad de análisis integrada tanto en el paradigma —lexicón—, como en el contexto sintagmático —corpus—.

Pero, para emprender el proceso de anotación morfológica se hace necesario observar no solo estos fundamentos de diseño de índole teórica, sino también otros de base computacional que salvaguarden la viabilidad en la codificación y el tratamiento y gestión eficaz de los datos. Será necesario, pues, formalizar todo ese conocimiento lingüístico que se pretende adjuntar a cada etiqueta, porque *solo lo formalizable puede ser programable* (Periñán, 2012:29).

Para estandarizar *lo formalizable*, los miembros del grupo EAGLES[32] presentaron en 1996 unas recomendaciones descriptivas para la *Anotación morfosintáctica de lexicones y corpora para las lenguas europeas*, con la intención de crear un marco común de anotación sintáctica que mantenga ciertos estándares garantizando la homogeneidad, consistencia y legibilidad entre aplicaciones de diferentes lenguas, esto es, permitiendo la reutilización de recursos de lo que hemos llamado *industrias de la lengua*. Estas recomendaciones de EAGLES son indicaciones muy generales que cada lengua deberá adaptar para la descripción de sus propias particularidades.

El estándar que propone este grupo de expertos (Monachini y Calzolari, 1996) se basa en el estudio comparativo de lexicones ya construidos y de corpus ya anotados, por lo que el resultado del trabajo es más una síntesis de los trabajos preexistentes que una propuesta original. Su objetivo será proponer una

[32] EAGLES (*The Expert Advisory Group on Language Engineering Standards*) es un proyecto financiado por la Comisión europea cuyo cometido es el de proporcionar unas directrices para la estandarización de las tecnologías lingüísticas. El objetivo de este grupo es promover la creación de estándares para el desarrollo, explotación y evaluación de recursos lingüísticos a gran escala. La definición de especificaciones y directrices se llevan a cabo en los distintos grupos de trabajo. Estos trabajan en las siguientes áreas: corpus textuales, lexicones computacionales, formalismos gramaticales, evaluación y tasación, y corpus orales.
El segundo grupo, dedicado a los lexicones, se encargó de analizar las principales prácticas de codificación lexicográfica, comparando recursos léxicos computacionales ya existentes en lenguas europeas como el alemán, catalán, danés, español, francés, griego, holandés, inglés, irlandés, italiano, portugués y sueco.

codificación consensuada que permita la comparación, *mappability* entre diferentes recursos —habidos y por haber— de las distintas lenguas del marco europeo (Civit, 2003:51). Para ello, ofrecieron un conjunto común de posibles distinciones morfosintácticas que podían y debían codificarse y que son las que se ofrecen como estándar de anotación[33].

En el esquema que proponen se establecen tres niveles de codificación que quedan recogidos, sintéticamente, en el siguiente cuadro:

RASGOS OBLIGATORIOS (L0)	Categoría de la palabra, *pos*: Nombre Verbo Adjetivo Pronombre Determinante Artículo Adverbio Adposición Conjunción Numeral Interjección Residual (esta etiqueta recoge barbarismos, abreviaturas, acrónimos, inclasificados, etc.)
RASGOS RECOMENDADOS (L1)	Núcleo mínimo de rasgos que deben codificarse: Flexión de género, número, persona, caso, rasgos semánticos —posesivos, demostrativos, indefinidos—, finitud y definitud, etc.
RASGOS OPCIONALES (L2)	Rasgos que no suelen codificarse: (Encajan solo en la descripción de algunas lenguas) Particularidades flexivas de género y número (común, invariable), *politeness* (en el caso de los pronombres de nuestra lengua), etc.

Esta propuesta admite varias soluciones para un mismo problema; para la anotación de los demostrativos, por poner un caso, se ofrecen dos posibilidades de categorización; bien como adjetivos del tipo determinativo, bien como determinantes. Esta misma propuesta de clasificación será válida para el resto de determinantes.

[33] Otros estándares de codificación utilizados en el ámbito de la anotación morfosintáctica son el TEI (*Text Encoding Initiative*) que fue el estándar seguido por el *British National Corpus*, el NERC (*Network of European Reference Corpus*) *y* el CES (*Corpus Encoding Standard Encoding Initiative*). Para una exposición detallada de estos y otros estándares, puede consultarse el trabajo de Martín de Santa Olalla (1999), en el que se pone en práctica el primero.

El documento de EAGLES, no obstante, no ofrece un método o criterio de discriminación para asignar las categorías a las palabras de la lengua; la propuesta se limita a recoger un esquema para catalogar la información seleccionada. Tampoco entrará en consideraciones propias de cada lengua, como la duplicación o doble catalogación de algunas palabras homónimas.

Por otro lado, deben tenerse en cuenta otras cuestiones de índole práctica; método y técnicas de anotación, diseño de la base de datos, nivel de detalle al que se pretende llegar, etc. La situación ideal es la de fundamentar lingüísticamente el conjunto de etiquetas, pero en la práctica se dan ocasiones en las que una buena fundamentación lingüística es inaplicable en términos de automatización. En estos casos se hace necesaria una solución de compromiso entre lo lingüísticamente deseable y lo computacionalmente viable.

El Capítulo siguiente se dedicará al desarrollo de una propuesta de etiquetario para el español que recoge las orientaciones de EAGLES y las concreta[34]. Se ofrecerá una descripción de algunos aspectos del Lexicón TIP. Se abordarán ciertos rasgos de diseño, buena parte de los aspectos teóricos que fundamentan su esquema de anotación y el estudio y tratamiento específico que se ha desarrollado para algunas piezas concretas. Este trabajo de concreción, revisión y anotación manual de piezas léxicas ha sido necesario para garantizar la precisión y rigor de la información codificada en los patrones de error que dan contenido a PatErr.

4.2 Los corpus

Como se expuso en el Capítulo 2, la Lingüística de Corpus es una disciplina que se dedica a la creación y explotación de los recursos lingüísticos empleados para la LC. Estos recursos representan un elemento ineludible para el desarrollo de las tecnologías del lenguaje, sus programas o aplicaciones. Junto con los lexicones computacionales o bases de datos léxicas se presentan los corpus, entendidos

[34] Otros ejemplos de etiquetario elaborados para nuestra lengua que siguen estos estándares son el de LEXCREA, (Sánchez León *et al.,* 1999) y el de LexEsp (Sebastián Gallés *et al*., 2000).

como un *conjunto estructurado de textos que forman una muestra representativa de uso real de la lengua*.

Aunque en la actualidad el concepto de corpus ha cambiado mucho con respecto al que manejaban los primeros lingüistas que lo empleaban como recurso para sus investigaciones, su finalidad última es siempre la misma; entender mejor cómo funciona el lenguaje humano. El objetivo que motive su uso puede ser diverso; obtener datos para una investigación teórica, para un programa de traducción automática, para confeccionar un diccionario o para construir un conversor de texto a habla.

Estos repertorios de lengua escrita o hablada son una herramienta esencial en el entorno computacional; proporcionan las bases necesarias para analizar la lengua y determinar sus características, son la base de entrenamiento y aprendizaje de ciertos automatismos, permiten verificar empíricamente una teoría gramatical y, por último, son bancos de pruebas de técnicas o aplicaciones de ingeniería lingüística que ayudan a determinar el buen funcionamiento y la cobertura de los programas. Dicho de otro modo, la función principal de un corpus, tanto textual como oral, es establecer la relación entre la teoría y los datos; el corpus tiene que mostrar a pequeña escala cómo funciona un lenguaje natural. Para ello es necesario que esté diseñado correctamente sobre unas bases estadísticas apropiadas que eviten sesgos y aseguren que el resultado sea efectivamente una muestra fiel de la realidad (Torruella y Llisterri, 1999:46).

Pero cualquier colección de materiales no da pie por sí mismo a la constitución de un corpus en el contexto del PLN. Este ha de cumplir una serie de requisitos; un diseño coherente, presencia de marcas que definan su estructura según unos estándares comúnmente aceptados y una documentación completa que permita conocer la procedencia y características del material que conforma el corpus. Según Sinclair, un corpus en el sentido moderno será:

> [...] a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language (1996:4).

Aunque según esta definición los criterios o herramientas informáticas quedan al margen de los límites del objeto, resulta inevitable, en el actual estado de

las cosas, la presencia de estas tecnologías para la construcción de un corpus. Asumiendo esta realidad, Sinclair define el concepto de *corpus informatizado*,

> [...] a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance (1996:4).

Con todo y con esto, los *corpus crudos*, esto es, los corpus que contienen exclusivamente las palabras y signos de los textos originales, tienen una utilidad bastante limitada para la investigación lingüística con corpus. La necesidad de completar los datos almacenados con algún tipo de información lingüística motiva la aparición de *corpus anotados,* que se presentan codificados y etiquetados con criterios explícitos y de manera homogénea para diversas tareas relacionadas con la recuperación de la información.

El lingüista que trabaja con corpus realiza una consulta sobre un conjunto de datos y recibe como resultado de una búsqueda "ciega" llevada a cabo por la máquina todos los casos —a veces en cantidades realmente avasalladoras— que responden formalmente a lo que se le ha sido solicitado. Ante este panorama de cantidad ingente de datos en bruto, solo una codificación, ordenación y organización de la información pueden salvarnos del naufragio en un mar inmenso de datos (Torruella y Llisterri, 1999:45).

Estos procesos de codificación, catalogación y anotación sirven, pues, para poder acotar y modelizar las búsquedas utilizando herramientas de filtro que aporten luz a la ceguera de esas búsquedas automáticas. Son, a la postre, los instrumentos que nos permiten acceder a los datos para, a partir de nuestra interpretación, convertirlos en información.

Como puede intuirse, las ventajas que ofrecen los corpus anotados son múltiples. Por un lado, facilitan la búsqueda de información porque permiten localizar de forma rápida una palabra, secuencia de palabras o incluso una categoría o esquema gramatical en décimas de segundo. Por otro, permiten acotar e identificar todos los casos en los que una palabra o secuencia de palabras aparece en el corpus incluyendo, habitualmente, su cotexto inmediato anterior y posterior del que pueden extraerse sus concordancias (Villayandre, 2008).

Buscando apoyar la construcción de corpus, la iniciativa EAGLES, como en el caso de los lexicones, emite unas recomendaciones básicas para la construcción de corpus útiles para el PLN que básicamente recogen las orientaciones desarrolladas por McEnery y Wilson (1996 y 2001) presentadas en el Capítulo 2, en el epígrafe dedicado a la Lingüística de Corpus[35].

A continuación, se ofrece una breve reseña sobre los procesos por los que suele pasar la CONSTRUCCIÓN DE UN CORPUS dedicado a tareas de PLN.

4.2.1 La codificación

Se trata de un proceso cuyo destino es la introducción de marcas que definan la estructura del corpus. Existen varios estándares para el marcaje de textos como el TEI (*Text Encoding Initative*), basados en el uso de SGML (*Standard Generalized Markup Language*) y más recientemente XML (*Extensible Markup Language*).

La información que se añade al texto, en este caso, no es de tipo lingüístico sino estructural; señala los títulos, subtítulos, la división en párrafos, etc., con un procedimiento análogo al que encontramos en el lenguaje HTML (*Hyper Text Markup Language*), estándar utilizado para la codificación de los textos que se publican en forma de página web.

4.2.2 La anotación

En cuanto a la anotación, que adjunta a cada unidad discriminada información estrictamente lingüística, será necesario explicitar su esquema de anotación o etiquetario, entendido este como un conjunto de convenciones por las que se asocia cada palabra del corpus a unos rasgos de información gramatical y

[35] Según las recomendaciones de EAGLES, un corpus deberá observar las siguientes consideraciones:
1) El corpus debe ser lo más extenso posible de acuerdo con las tecnologías disponibles en cada época.
2) Debe incluir ejemplos de amplia gama de materiales en función de ser lo más representativo posible
3) Debe existir una clasificación intermedia en los géneros entre el corpus en total y las muestras individuales.
4) Las muestras deben ser de tamaños similares.
5) El corpus, como un todo, debe tener una procedencia clara.

semántica, según sea el caso, que pueda ser relevante para su análisis una vez que se ha extraído de texto (Llisterri y Moure, 1996:176).

Deberá, además, estar estructurado sobre una base de datos dotada de un sistema de interrogación que permita la recuperación de la información textual en cada consulta a la que se le someta. De este modo se posibilita que un sistema automático de procesamiento del lenguaje pueda interpretar la información lingüística y explotar todas sus posibilidades.

Los niveles de anotación y las etapas de tratamiento e interpretación de la lengua que se consideran para los corpus escritos se asimilan a los tipos de información susceptible de ser anotada en un lexicón.

Independientemente de cuál sea el nivel de anotación, este proceso puede tener tres ejecuciones diferentes:

- *Automática*. Un programa informático anota el corpus según las reglas y algoritmos que han sido previamente programados o aprendidos por el programa ejecutor[36].
- *Manual*. En algunos casos, es precisa la intervención total y exclusiva de expertos para esta tarea de anotación, bien sea porque no existen las herramientas informáticas para ello, bien porque los porcentajes de error que arrojan los programas de anotación son muy altos.
- *Semiautomático*. En ocasiones, los resultados que ofrecen los programas informáticos no son fiables o exactos, por lo que exigen la intervención de analistas humanos expertos —post-edición—. Primero actúa el programa, posteriormente los lingüistas revisan los resultados aportando mayor fiabilidad en los resultados.

Como en el caso de los lexicones, con independencia de cuál sea la técnica empleada para su desarrollo, su diseño debe garantizar que el producto sea *actualizable* y *reutilizable*, conceptos cruciales que se asocian a la vigencia y

[36] Para que una herramienta de anotación automática realice su tarea de un modo eficaz, se requiere una etapa de entrenamiento con un corpus anotado manualmente. Por otra parte, las anotaciones automáticas requieren una revisión manual a cargo del experto para mejorar el analizador, refinando sus reglas e incorporando conocimientos que se aplicarán a la anotación de nuevos corpus.

rentabilidad del corpus, es decir, a su capacidad para ser sostén de diversas aplicaciones con diferentes medios y fines (Torruella y Llisterri, 1999:46).

4.2.3 Corpus TIP

El Corpus TIP nace de un proyecto de ingeniería informática de este grupo de investigación. El objetivo de este proyecto fue crear un programa capaz de rastrear y recapitular miles de portales de internet en español (nacional e internacional) con el fin de extraer párrafos únicos, que posteriormente serán segmentados en frases. Estos datos deben dar contenido a un *corpus electrónico de referencia* del español panhispánico actual escrito en páginas web de internet.

El corpus, que actualmente alberga 100 millones de párrafos diferentes —alrededor de 3 300 millones de palabras— puede ser anotado parcialmente por programas creados *ad hoc* a partir de la reutilización de recursos desarrollados en el seno de este grupo de investigación; un lematizador, un flexionador y un conjugador[37].

El Corpus TIP ha sido un recurso esencial para la investigación lingüística previa a la codificación de los patrones que constituyen el repositorio de errores PatErr. Los datos que ofrece el corpus, tras una tarea de exégesis por parte del lingüista, se convierten en información relevante que debe ser formalizada en términos computacionales. Por otro lado, el corpus ha permitido corroborar los *aprioris* derivados de nuestra competencia lingüística, revelando, mediante datos,

[37] Las cifras alcanzadas en cuanto a volumen de datos marcan, como en el caso del Lexicón, una diferencia abismal entre el Corpus TIP y otros corpus. Una de las colecciones en lengua inglesa con más prestigio por su utilidad es el *British National Corpus,* que alcanza los cien millones de palabras del inglés moderno, tanto escrito como hablado (puede consultarse en http://www.natcorp.ox.ac.uk/). Otra de las obras relevantes para la lengua inglesa es el *Bank of English* (1991, COBUILD Corpus) corpus monitor, abierto, enfocado a la creación de diccionarios que se ve incrementado constantemente. Con todo, actualmente no supera los 650 millones de palabras.
En el contexto de nuestra lengua, cabe insistir en el valioso trabajo de Mark Davies, corpus anotado que recoge textos del español histórico y moderno que alcanza los cien millones de palabras (http://www.corpusdelespanol.org/). El CREA y el CORDE, corpus elaborados por la RAE, alcanzan entre los dos los 270 millones de palabras. Están codificados y presentan ciertas anotaciones (generalmente de carácter extralingüístico), pero carecen de anotación morfosintáctica, lo que limita enormemente su explotación en el marco de la LC, no así en otras investigaciones de carácter lingüístico. (http://corpus.rae.es/creanet.html y http://corpus.rae.es/cordenet.html).

el funcionamiento regular de nuestra lengua, sus excepciones y las especificaciones necesarias para llevar a cabo su tratamiento automático.

4.3 Las herramientas de análisis lingüístico

Los corpus por sí solos no son suficientes para facilitarnos datos exhaustivos sobre el comportamiento del lenguaje. Para poder aprovechar al máximo las informaciones que contienen es necesario disponer de herramientas adecuadas para su procesamiento y explotación. En este sentido hay que decir que tan importante es el corpus como las herramientas de análisis lingüístico que actúen sobre él.

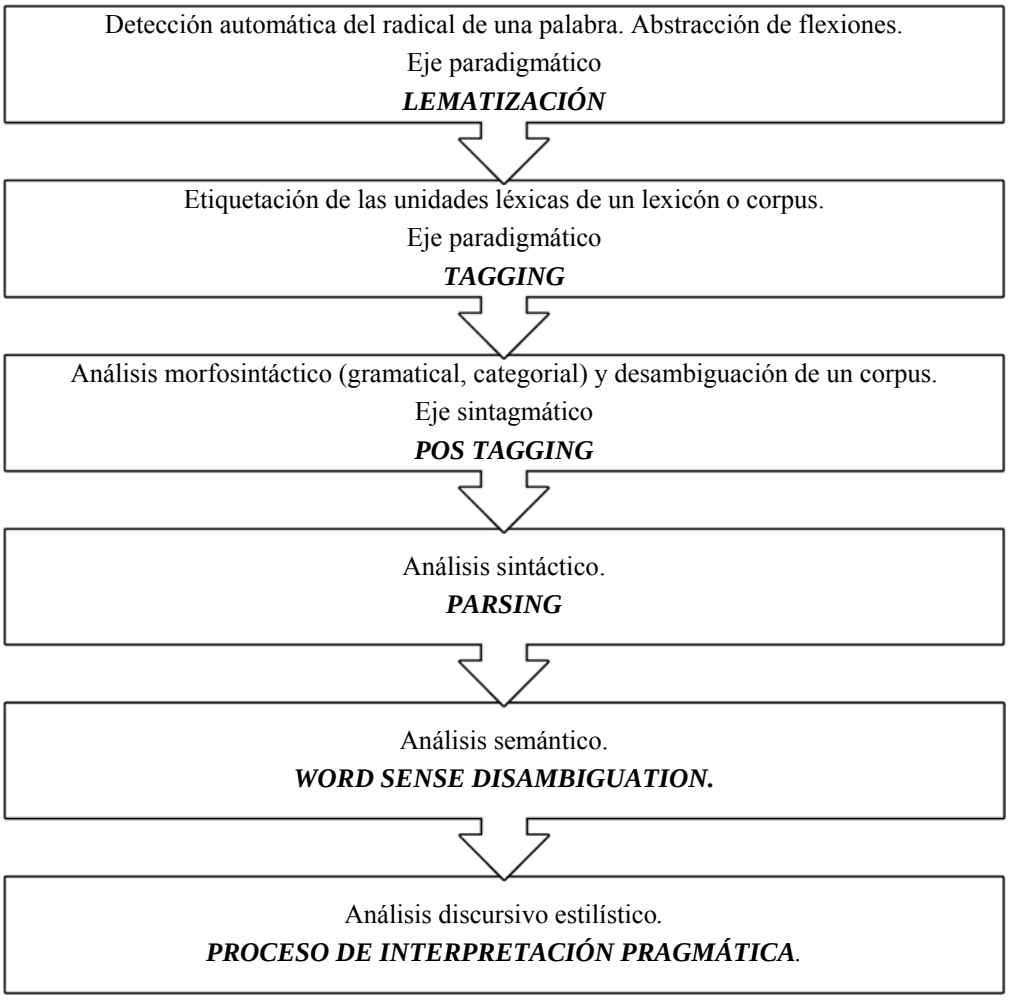
Por otro lado, la inmensa mayoría de los sistemas y productos generados por las industrias de la lengua escrita necesitan herramientas que proyecten análisis sobre alguno —o todos— los estratos de la lengua. Las tecnologías que sustentan esos programas incluyen alguna forma de tratamiento de la lengua y requieren la actuación de una serie de procesadores lingüísticos que actúen de forma estratificada: nivel morfológico, nivel léxico, nivel sintáctico, nivel semántico, nivel pragmático-discursivo, etc.

Estos procesadores locales pueden funcionar de forma aislada o colaborativa y, en palabras de Llisterri (2003:21) su acción puede sintetizarse de la siguiente manera:

> [...] realizan de un modo automático las operaciones que un especialista en morfología y sintaxis conoce a la perfección: extraer la raíz de una palabra, segmentar la palabra en morfemas, asignarles la categoría gramatical correspondiente, determinar la parte de la oración a la que pertenece la palabra, y descomponer una frase en sus constituyentes indicando la función sintáctica de cada uno de ellos.

La siguiente tabla de proceso, expresada en términos comúnmente utilizados en LC, permite ilustrar la sucesión de estas etapas[38]:

[38] Los corpus orales tienen además niveles que exige el discurso oral, como la transcripción fonética o fonémica, la anotación prosódica, el tratamiento de ruidos que intervienen en la comunicación.



Aunque, como es lógico, la complejidad del tratamiento lingüístico depende de la complejidad de la aplicación para la que trabaja, hay procesos iniciales como la segmentación, la lematización y la desambiguación que suelen aparecer en prácticamente todos los sistemas.

Como se señaló al inicio de este capítulo, los recursos de análisis lingüístico que han desarrollado en el grupo de investigación CLTIP han sido esenciales en el planteamiento que se ha llevado a cabo para la constitución del compendio de errores que aquí se propone. A continuación, se expone una breve descripción de las herramientas que han sido necesarias para la concreción de esta propuesta.

Para un análisis ampliado acerca de estos niveles, procedimientos, herramientas y terminología específicos de LC, puede consultarse la siguiente bibliografía general: Leech (1993, 2005), McEnery y Wilson (1996, 2001), Wynne (2005).

4.3.1 Un lematizador

La lematización es un proceso de agrupación de todas las variantes de una unidad léxica. Para llevar a cabo este procedimiento es necesario recuperar la morfología de las palabras, es decir, recorrer en sentido inverso el mecanismo de formación que ha sufrido la palabra hasta llegar a su lema o forma canónica, que representa todas las posibilidades que le ofrece esa forma a la lengua.

Un lematizador, por lo tanto, es una herramienta asociativa que vincula una forma derivada a su forma origen, desprovista de todos los afijos que no formen parte de la raíz.

Como puede intuirse, la lematización es un proceso que ha aportado una mayor flexibilidad en las búsquedas dentro de los textos o diccionarios. Aplicando estos programas, dotamos de cierta intuición lingüística a las herramientas de PLN. Hasta hace no mucho tiempo, el DRAE electrónico no contaba con un lematizador que explicitara estas relaciones entre palabras. Esta carencia obligaba al sistema a fracasar ante una palabra flexionada o conjugada como *balconcito* o *caminando,* y ofrecía la siguiente notificación.

***Aviso**: La palabra balconcito no está en el Diccionario*[39].

La incorporación de esta herramienta en el lexicón que sirve de base para el DRAE permite, en la actualidad, llevar a cabo búsquedas de palabras flexionadas o conjugadas que el sistema relacionará con los lemas de los que derivan.

Para concretar el planteamiento del que deriva el recurso que presentamos se ha contado con el Flexionador TIP, programa que se nutre del lematizador TIP para llevar a cabo sus ejecuciones[40].

[39] Actualmente, ante una búsqueda como la propuesta, la información que ofrece el sistema es la siguiente: "La palabra balconcito no está registrada en el Diccionario. La entrada que se muestra a continuación podría estar relacionada: *balcón*".

[40] Como este programa que puede revisarse en http://tip.iatext.ulpgc.es/flexionador/ se han desarrollado otros programas similares para el español. Pueden encontrarse demostraciones del funcionamiento de estos sistemas en las páginas web de varios grupos de investigación universitaria;

- CLiC; http://clic.ub.edu/en/node/37
- Molino de Ideas: http://www.molinolabs.com/lematizador.html
- Signum: http://lenguaje.com/herramientas/lematizador.php

4.3.2 Un flexionador

El siguiente hito en el procesamiento del lenguaje supone un análisis morfológico completo, en el que un programa automático de marcaje, conocido en este ámbito como *tagger,* adhiere a cada forma una etiqueta con su categoría gramatical y circunstancias flexivas.

Se trata de un procedimiento de análisis superficial y presintáctico que básicamente consiste en la descomposición de la palabra dada en un manojo de fragmentos/monemas: raíz, prefijos, sufijos, tanto flexivos como derivativos, y en algunas lenguas como la nuestra, infijos. Todos estos morfemas compositivos forman parte de conjuntos finitos y actúan siguiendo determinadas reglas de combinación.

El tipo de información que suele proveer un *tagger* como el Flexionador TIP[41] se funda en los siguientes criterios:

- *Categorías gramaticales válidas para la unidad tratada*
  En los casos en los que la forma presente ambigüedad, esto es, pueda desarrollar más de una categoría gramatical, el *tagger* debe proporcionar tantas etiquetas como categorías puedan otorgársele a esa forma. Así, un flexionador solvente deberá ofrecer tres etiquetas diferentes para la forma ambigua *venda;* dos verbales —*vender* y *vendar*— y una nominal.

- *Rasgos morfológicos*
  Rasgos lingüísticos como el género, el número, la persona, la flexión verbal, el grado en el caso del adjetivo, las flexiones derivativas, etc. son dosis de información relevante que habitualmente ofrecen los flexionadores.

- *Información de naturaleza semántica*
  Hay flexionadores, como el utilizado en este trabajo, que ofrecen prestaciones que exceden el ámbito de la morfología y proporcionan

[41] http://tip.dis.ulpgc.es/flexionador/ Santana *et al,* (2004, 2003).

> al usuario información semántica mediante etiquetas como *numeral, cuantificador, indefinido* o, en el caso de los verbos *transitivo, impersonal, pronominal,* etc.

Una de las posibles explotaciones de PatErr requiere, como se señaló, la etiquetación en términos categoriales del texto objeto de la revisión. Esta tarea de etiquetación de cada elemento en el texto recaerá sobre el flexionador que, apoyado en el lematizador, deberá rastrear el origen de la forma, captar su lema y adscribirle su flexión y la categoría gramatical.

4.3.3 Un conjugador

El objeto de este tipo de programas es el desarrollo automático de todo el repertorio de formas conjugadas que puede ofrecer un verbo. Son múltiples los conjugadores que se han desarrollado en el entorno del PLN y están disponibles en la red, aunque la cobertura, la fiabilidad y las prestaciones que presentan difieren de una aplicación a otra.

Para el desarrollo y explotación de PatErr es necesaria la concurrencia, junto con el Lematizador, el Flexionador y el Lexicón TIP, de un conjugador[42]. El que se ha utilizado se nutre de una base de datos de más de 14 000 verbos previamente conjugados e incorpora las diferentes conjugaciones aceptadas por la Asociación de Academias de la Lengua Española; la opción rioplatense de *vos* y las variantes de respeto *usted* y *ustedes*.

Con independencia de los algoritmos y técnicas que permitan ejecutar las herramientas de análisis que se han expuesto, lo que es evidente es que estos programas requieren surtirse, en mayor o menor medida, de conocimiento lingüístico formalizado. De algún modo realizan de forma automática las operaciones de segmentación y análisis que llevaría a cabo un experto en morfología y en sintaxis enfrentado a la misma tarea. La presencia del lingüista es, por lo tanto, imprescindible en el desarrollo de las herramientas de análisis de

[42] Carreras *et al.* (2011). http://tip.iatext.ulpgc.es/conjugador/

PLN, en una primera fase para definir las etiquetas que se asignan en el análisis y, en las etapas sucesivas, para validar y refinar los resultados de su ejecución (Llisterri, 2003).

Capítulo 5

Propuesta de etiquetario para el español. El Lexicón TIP

Un lexicón que pretenda ser realista y eficaz en términos computacionales debe plantearse y diseñarse como un recurso dinámico en constante crecimiento. Por otro lado, deberá ofrecer, junto a las formas que lo constituyan, la información relativa a las posibilidades que cada una presenta en el discurso. El lexicón deberá captar la información de cada uno de sus elementos ya se inscriban en el paradigma, ya se actualicen en el sintagma. De la compilación, clasificación y anotación exhaustiva de todos estos rasgos informativos y matices que ofrecen las palabras, dependerá el éxito y la fineza del procesamiento automático de la lengua natural a la que se dedique el lexicón.

5.1 Contenido

El contenido que da cuerpo al Lexicón TIP se constituye a partir de las UNIDADES LÉXICAS compuestas por una única palabra, es decir, los grafemas *separados por espacios en blanco*[43].

En términos generales, las unidades léxicas del español coinciden biunívocamente con una única palabra (NGLE §1.3e). No obstante, debe advertirse que, en el caso de las *multipalabras* —más de una palabra ortográfica se corresponde con una unidad léxica— y de las *contracciones*, este criterio ortográfico no coincide con el morfosintáctico. Es por esta razón por la que, en este ámbito, prefiere hablarse de unidades léxicas en lugar de palabras. Estarán presentes tanto las formas canónicas, como las conjugadas, flexionadas y derivadas. Se incluyen, además,

[43] Las unidades léxicas compuestas por más de una palabra podrán ser identificadas, tratadas y recopiladas en trabajos posteriores a partir de la técnica de N-gramas que, como se vio en el Capítulo 3, permite tener en cuenta parte del cotexto local en que se inscriben las palabras en la lengua, en el eje sintagmático. Esta técnica se fundamenta en la asunción de que la probabilidad de aparición de una palabra está condicionada por las unidades contiguas.

nombres propios, apellidos y topónimos, abreviaturas y siglas, así como algunos extranjerismos.

Como se verá, son frecuentes los casos en los que una misma forma como *mucho* puede desarrollar varias categorías en la frase —adjetivo, pronombre, adverbio—. Según el diseño propuesto para nuestra base de datos, continente del lexicón, cada forma deberá presentar una única etiqueta identificativa en la que se exprese, entre otras, esta información categorial. Teniendo en cuenta esta premisa, una forma deberá tener tantas entradas —y etiquetas— como categorías pueda desarrollar en el discurso, de modo que se pueda garantizar la máxima cobertura sobre todas las posibilidades que presenta la lengua[44].

Según las buenas prácticas de diseño de base de datos podría aunarse en una única entrada las grafías similares y relacionarlas con las distintas categorías que pudieran desarrollar, pero este diseño es incompatible con el hecho de que las distintas categorías gramaticales conllevan, en la mayoría de los casos, algunos tipos de flexión y/o de derivación diferentes; *amanecer,* como verbo, presenta un repertorio de formas conjugadas relacionadas en la base de datos, mientras que el sustantivo *amanecer* se asocia a formas flexivas nominales.

5.2 Fuentes de extracción y su tratamiento

El Lexicón TIP surge a partir de un catálogo amplio de fuentes lexicográficas recogidas en diccionarios electrónicos o digitalizados. Mediante técnicas automáticas —algoritmos— se han rastreado y compilado todas las entradas contenidas en los siguientes diccionarios:

- Clave: Diccionario de Uso del Español Actual. Madrid: S.M.
- Versión electrónica de la edición de 1997. Consultado en http://clave.smdiccionarios.com/app.php.
- *Diccionario de la Real Academia Española*. Real Academia Española, Espasa Calpe.

[44] La polisemia no será, de momento, un factor que deba tenerse en cuenta en el diseño, dado que afecta a análisis semánticos que se escapan a los planteamientos y proyectos que se desarrollan actualmente en este grupo de investigación.

- Versión electrónica de la vigésima segunda edición, en proceso de enmienda. Consultado en http://www.rae.es/rae.html
- *Diccionario de Uso del Español de María Moliner*. Madrid: Gredos. Versión digitalizada de la edición de 1996.
- *Diccionario de voces de uso actual*. Alvar Ezquerra, M. Madrid: Arco Libros. Versión digitalizada de la edición de 1994.
- *Diccionario General de la Lengua Española Vox*. Barcelona: Biblograf. Versión digitalizada de la edición de 1997.
- *Gran Diccionario de la Lengua Española Larousse.* Barcelona: Planeta. Versión digitalizada de la edición de 1997.
- *Gran Diccionario de Sinónimos y Antónimos*. Madrid: Espasa Calpe. Versión digitalizada de la edición de 1991.

Tras este proceso de compilación de datos es necesario formalizar y homogeneizar la catalogación y el etiquetado de todos los ítems recogidos, así como la información relevante que quiera registrarse para su posterior uso computacional.

La metodología elegida para la creación de esta base de datos léxica es mixta; por un lado, se ha llevado a cabo una reutilización automática de recursos ya existentes, y por otro, se ha culminado la colecta de datos con otro proceso manual que ha permitido enriquecer el contenido y la cobertura del lexicón.

Es evidente que, mientras el lenguaje es un objeto dinámico en constante evolución, los diccionarios son, por definición, objetos estáticos. El lapso de tiempo que transcurre entre el proceso de compilación y la edición, publicación, digitalización y distribución de un diccionario, hace imposible que este pueda ser un reflejo vigente de una lengua, situación que se agrava cuanto más tiempo pasa desde su publicación.

Para subsanar esta realidad, y con el objetivo en el horizonte de ofrecer un lexicón realista, se ha hecho necesaria la inclusión manual de nuevos ítems que se han extraído a partir de un estudio que se hecho sobre el Corpus TIP. Del contraste de este corpus con el lexicón derivado de los diccionarios listados, se comprobó que una ingente cantidad de entradas léxicas no tenían réplica en el lexicón. Como cabe esperar, tras una revisión minuciosa de las 4 000 palabras no reconocidas que

aparecen en el corpus con mayor frecuencia se comprobó que una parte importante de estas entradas se corresponden con errores de ortografía, errores derivados de la escritura digital o errores de segmentación de palabras. No obstante, esta revisión manual nos ha permitido incrementar y actualizar el Lexicón TIP con más de 220 entradas nuevas que se justifican por su frecuencia de uso[45].

5.3 Cobertura. El lexicón en cifras

El Lexicón TIP almacena un listado de más de seis millones de palabras del español panhispánico anotadas, de las cuales 4 345 689 son distintas. Cuenta con 258 397 formas canónicas y de estas, 226 026 son entradas únicas, es decir, diferentes.

Estas cifras destacan con rotundidad si hacemos una comparación con otros lexicones anotados de nuestra lengua; solo LexEsp, *Léxico informatizado del español* se acerca a nuestras cifras, alcanzando los cinco millones y medio de palabras anotadas morfológicamente de modo automático (Sebastián Gallés *et al.*, 2000). El reparto de estas formas en términos de categorías gramaticales es:

Categoría gramatical	Formas canónicas
Sustantivos	119 353
Adjetivos	44 322
Adverbios	71 594
Verbos	21 707
Pronombres	668
Conjunciones	58
Preposiciones	54
Otras	734

[45] Aunque la cantidad significativa de las nuevas incorporaciones se corresponde con la categoría de los adjetivos, gran parte de los elencos de categorías abiertas han sido incrementados. La inclusión de estso términos, por un lado, harán del lexicón un producto actualizado, y por otro, evitarán colapsos en etapas posteriores de procesamiento de textos.
SUSTANTIVOS: *talibán*, *selfie*, *chat*, *coach*, *fibromialgia*, *píxel*, *tofu*.
ADJETIVOS: *investigativo*, *gore*, *colaborativo*, *friki*, *ambientalista*, *indie*, *vegano*.
VERBOS: *empoderar*, *esponsorizar*. *monitorear*, *reseterar*.
SIGLAS-ACRÓNIMOS: *USB*, *ONG*, *DVD*, *ADSL*
MORFEMAS DERIVATIVOS: *ciber-*, *bio-*, *multi-*, *-cardio-*.
Por otro lado, se han incluido un buen número de marcas tecnológicas como *Google*, *Facebook*, *Twitter*, *Instagram*, *PhotoShop*, *WhatsApp*, etc.

Palabras	6 288 673
Palabras únicas	4 346 519

Cada forma canónica tiene asociadas todas las flexiones que le corresponden, con independencia de su frecuencia de uso en el español. Las flexiones consideradas para el Lexicón TIP son las siguientes: género, número, superlativo, diminutivo, aumentativo y despectivo.

A partir de este contenido, la cobertura que actualmente presenta el lexicón sobre un corpus de 3 366 millones de palabras extraídas de internet es del 99,65%, es decir, 12 millones de formas no están recogidas en Lexicón TIP[46].

5.4 Información compilada sobre los componentes

Como se ha sugerido, cada entrada del lexicón tiene asociada una etiqueta con información lingüística que concierne a varios niveles de análisis lingüístico: el género, el número, la flexión —en caso de que desarrolle alguna—, el número de sílabas, la posición de la sílaba tónica, el número de etimologías, que dará lugar a distintas entradas con una misma grafía[47], el número de acepciones totales que posee en el DRAE, la frecuencia de aparición en el CREA, una indicación de si es antigua o desusada y, por último, la categoría gramatical.

Los asuntos relacionados con el significado no pasarán totalmente desapercibidos en este lexicón pues sobre la base del DRAE se han extraído ciertas características conceptuales de cada palabra, como el número de acepciones que posee o sus etimologías.

La siguiente tabla ofrece todos los campos de información, el esquema de anotación, que se aportan para cada forma canónica en el Lexicón.

[46] Una descripción pormenorizada de estos datos puede consultarse en Carreras *et al.* (2012), y en http://tip.dis.ulpgc.es/es/corpus-lexico-tip?showall=1.

[47] La etimología es un factor crucial para tomar en cuenta a la hora de diseñar la base de datos, pues afecta a la derivación, la sinonimia, y otras características de la lengua.

PALABRAS | CATEGORIAS_GRAMATICALES | FORMAS_CANONICAS

Nombre del campo	Tipo de datos	Descripción (opcional)
IDFORMA_CANONICA	Autonumeración	Clave primaria de la forma canónica.
FORMA_CANONICA	Texto corto	Forma canónica.
IDCATEGORIA_GRAMATICAL	Número	Categoría gramatical de la forma canónica hasta 12-12-2012. Para ParamText y otras aplicaciones.
IDCATEGORIA_GRAMATICAL_G	Número	Categoría gramatical de la forma canónica que para los sustantivos está desglosada por género y número. Para el flexionador.
IDFLEXION	Número	Flexión de la forma canónica.
NUM_FLEXIONES_BASICAS	Número	Número de flexiones básicas (género y número) que tiene la forma canónica.
NUM_SILABAS	Número	Número de sílabas de la forma canónica.
SILABA_TONICA	Número	1 - aguda, 2 - llana, 3 o más esdrújula.
MAX_ETIMOLOGIA	Número	Número máximo de entradas diferentes en el DRAE por tener diferente etimología.
NUM_ETIMOLOGIA	Número	Orden de etimología en el DRAE
NUM_TOTAL_ACEPCION	Número	Número total de acepciones de la palabra con independencia de la categoría gramatical en el DRAE.
POS_ACEPCION	Número	Posición de la acepción de la primera aparición de esta categoría gramatical en el DRAE.
NUM_ACEPCION	Número	Cantidad de acepciones con esta categoría gramatical en el DRAE.
POS_ACEPCION_SEC	Número	Posición de la acepción de la primera aparición de esta categoría gramatical en el DRAE al final de párrafo.
NUM_ACEPCION_SEC	Número	Cantidad de acepciones con esta categoría gramatical en el DRAE al final de párrafo.
Frec_CREA	Número	Frecuencia de aparición en el corpus CREA de la forma canónica y todas sus flexiones.
Frec_Corpus	Número	Frecuencia de aparición en el corpusdelespanol.org de Mark Davis de la forma canónica y todas sus flexiones.
Antigua o desusada	Sí/No	Algunas antiguas o desusadas de número de acepciones 1 y número de entradas-etimologías 1
DRAE XXIII	Sí/No	Está en la última versión del DRAE
NGLE	Texto corto	

Algunas de estas informaciones que configuran los rasgos de la etiqueta de cada forma canónica y palabra, son el producto de un *proceso de anotación automática* llevada a cabo inicialmente sobre el lexicón. Este procedimiento, como ya se reseñó, se basa en la extracción, mediante algoritmos y programas creados *ad hoc,* de las dosis de información relevante que cada forma canónica presenta en las fuentes lexicográficas.

La información aportada por los diccionarios es escrupulosa y rica en cuanto a las definiciones nocionales de las formas, que es su objetivo principal, pero carece, especialmente en el caso de las categorías funcionales, de sistematicidad en términos gramaticales. No es extraño que unidades léxicas que tienen un comportamiento similar morfológica, sintáctica y/o semánticamente no reciban un tratamiento homogéneo en los diccionarios (Periñán, 2005). Esta falta de rigor gramatical en la anotación categorial de las palabras, queda patente, por poner un caso, en las disonancias teóricas que se observan entre las diferentes fuentes de la RAE (DRAE, NGLE, *Manual*, DPD)[48]. El inconveniente se acrecienta cuando se cotejan datos procedentes de fuentes distintas, que no siempre coinciden en cobertura y categorización[49].

[48] Debemos reconocer, no obstante, el esfuerzo continuo en forma de enmiendas y actualizaciones por parte de la Academia. El seguimiento diario de la versión digital del DRAE revela un decidido intento de homogeneización por parte de la institución. Este evidente impulso de revisión (en la 21ª y 22ª edición) surge tras el análisis de los datos empíricos ofrecidos por el corpus CREA, que guía la adición de artículos en el diccionario.

[49] Porto Dapena (2004) expone este problema de ausencia de criterios en los diccionarios para la lematización, que se acrecienta en casos de polimorfismo como el de los pronombres.

Por otro lado, estos automatismos dedicados a la extracción de información, a pesar de trabajar con unos bajos niveles de error, requieren la validación manual de los datos que resultan de los procesos de selección automática. Al respecto de estas tareas, sostiene Periñán:

> La revisión y corrección de errores de los datos extraídos automáticamente es una tarea costosa tanto en tiempo como en recursos humanos […] que puede ser superior al trabajo implicado en la construcción manual de la misma base de conocimiento (2005:14).

Una de las tareas de investigación pormenorizada que ha tenido que afrontarse para garantizar la consistencia y cobertura del contenido de los patrones de error codificados de PatErr ha sido, precisamente, el análisis exhaustivo y la catalogación de aquellas palabras pertenecientes a las llamadas *categorías gramaticales*[50] que se suelen oponer a las *categorías léxicas*. Las formas que pertenecen a estas categorías han sido analizadas tanto en el eje paradigmático, como en el sintagmático con el fin de desbrozar y esclarecer la identidad de estas controvertidas formas que resultan cruciales para la formalización de los patrones de error.

5.5 Concreción de la propuesta

El etiquetario que hemos diseñado para el español, y del que hemos partido para la catalogación y etiquetación del Lexicón TIP sigue, en lo esencial, las recomendaciones enunciadas por EAGLES aunque, como se verá, adapta la taxonomía para dar cabida a categorías que se asientan en diferentes niveles.

[50] En la lingüística moderna de corte generativista se prefiere, como ya veremos, el adjetivo *funcional*.

CATEGORÍAS EN EL LEXICÓN Gramaticales
artículos
adjetivos
sustantivos
pronombres
verbos
adverbios
preposiciones
conjunciones
conectores discursivos
interjecciones
onomatopeyas

CLASES TRANSVERSALES Semánticas
demostrativos
posesivos
cuantificadores
indefinidos
numerales
relativos
interrogativos
exclamativos

CATEGORÍAS EN EL TEXTO Discursivas
determinantes

En el Lexicón TIP —observando el nivel L0 de EAGLES de carácter obligatorio—, todas las formas presentarán, junto a su forma canónica, información acerca de su flexión —género, número, tiempo, caso y persona— cuando la manifieste, y categoría gramatical. En lo que respecta a los niveles L1 y L2 se codificarán, según nuestra propuesta, algunos de los aspectos semánticos y funcionales que recomienda EAGLES; información semántica en el caso de las clases transversales e información sintáctica-funcional en el caso de los determinantes y los verbos.

A continuación, se exponen los fundamentos teóricos de carácter lingüístico que justifican el diseño y desarrollo del etiquetario que se propone.

5.5.1 Fundamentación lingüística

5.5.1.1 *Preámbulo. Sobre las unidades de la gramática*

Uno de los cometidos fundamentales de la teoría gramatical es señalar las generalizaciones lingüísticamente significativas que expresan las regularidades

básicas de la lengua. Para ello, cualquier tratado gramatical debe partir de una delimitación previa de las unidades, niveles de análisis e instrumentos básicos que permitirán la descripción precisa de los fenómenos que pretenda abordar.

En cada *nivel* de la lengua se reconocen determinadas *unidades* que tienen su razón de ser dentro de la *teoría* en que se formulan, según el nivel en el que se asienten y con respecto a la relación que guardan con las demás unidades del mismo nivel. Estas unidades, que se recogen en las categorías, constituyen una de las nociones básicas de toda teoría sintáctica, un criterio operativo fundamental en la clasificación de los elementos de la gramática. Son, como sostiene Gutiérrez Ordóñez, *un aspecto básico en la doctrina gramatical de toda época* (1994:71).

Pero la clasificación unánime de estas unidades en categorías es, a pesar de la tinta y los esfuerzos concedidos, un desiderátum para la teoría gramatical. Con frecuencia, las clases presentan propiedades y límites que no son compartidos por todos los elementos que la integran. Por otro lado, resurge constantemente la dificultad de establecer una clara diferenciación en algunas unidades que parecen participar de varios paradigmas. Sirva como ejemplo el caso de formas como *hasta, excepto* o *cuando* que se debaten entre los títulos de *conjunción, preposición* y *adverbio*. Estas fluctuaciones entre categorías son procesos comunes en todas las lenguas y provocan, con frecuencia, que los perfiles intercategoriales se debiliten y aparezcan borrosos.

El término CATEGORÍAS GRAMATICALES se emplea con varios sentidos en la ciencia lingüística, de hecho, el mismo término *categoría* ha sido objeto de múltiples sinónimos, enfoques y definiciones. En la tradición gramatical, procedente de las gramáticas clásicas, era común el término *partes de la oración,* que hacía referencia a las clases o categorías en las que se agrupaban las palabras tomando como base propiedades de índole diversa; semánticas, morfológicas o flexivas y combinatorias o sintácticas.

Según la primera *Gramática* de la Academia (1771), *son nueve en nuestra lengua las partes de la oración*: nombre, pronombre, artículo, verbo, participio, adverbio, preposición, conjunción, interjección. […] *de suerte que cualquier palabra ha de ser precisamente, o nombre, o pronombre, o artículo, o verbo,* etc. (RAE, 1984:

120-121)[51]. Como se puede advertir, no hay cambios drásticos entre la clasificación de los gramáticos de entonces y las taxonomías que se manejan en la actualidad.

En el marco de la gramática tradicional[52], ha sido norma común respetar, en lo esencial, el sistema de clasificación de las palabras establecido en las gramáticas griega y latina. Únicamente se observan algunos retoques aquí o allá para adaptarse mejor a cada lengua debido a que, en ocasiones, no presentaban —superficialmente— ni el mismo número de elementos lingüísticos, ni su misma distribución.

A pesar de los reajustes y las redefiniciones hay algo que sorprendentemente no suelen cuestionarse los gramáticos;

> […] la pertinencia y el fundamento de dichas clasificaciones; no se pone en duda el principio que subyace a tales divisiones y que, a grandes rasgos, podemos decir que consiste en establecer paradigmas donde se integren todos los elementos que se reconocen diferentes de los otros y semejantes entre sí, […] (Varela, 1979:55).

La categorización lingüística, entendida como el proceso ordenador del que resultan las categorías —en cualquiera de las acepciones del término—, conllevará, por lo tanto, un proceso de necesaria estereotipación, que será de gran utilidad para obtener generalizaciones teóricas aunque, como señala Bosque, en ocasiones representen un obstáculo para profundizar en el análisis de algunos fenómenos (1979:81).

No obstante, la dificultad declarada de esta tarea no ha mermado el interés por estos temas en los estudios lingüísticos. El examen de los límites intercategoriales, de los movimientos lingüísticos como la *recategorización*[53], de las categorías híbridas y de los mismos criterios que subyacen a la categorización se han convertido en objeto de estudio para disciplinas como la morfología, la sintaxis, la semántica y la lexicografía.

---

[51] *Cf.* Alcaraz Varó y Martínez Linares (2004:432).

[52] Entendiendo como tal la gramática no estructuralista y no generativista.

[53] Hay dos fenómenos lingüísticos de suma frecuencia que llevan consigo un proceso de recategorización; la *sustantivación* de elementos no nominales y la *elipsis* de elementos nominales que, puede ser su causa en algunos contextos. Ambos fenómenos ponen en tela de juicio ciertos aspectos muy debatidos en el ámbito de la gramática teórica; por un lado, el cambio de categoría de una palabra —al menos a nivel funcional—, por otro, el supuesto poder sustantivador del artículo.

Algunos de los problemas terminológicos y conceptuales con los que se enfrenta la delimitación de las clases y la descripción de las unidades son, por un lado, la disparidad de los criterios empleados en la definición de las categorías gramaticales, con la consiguiente falta de acuerdo en clases como los determinantes, pronombres o adverbios, y por otro, el extendido recurso de utilizar los procesos de cambio de categoría, —en particular la sustantivación—, como mecanismos de análisis (Eguren y Soriano, 2006:11).

Mucho se ha teorizado desde todos los paradigmas lingüísticos acerca de estos aspectos, pero por falta de necesidad —nunca se ha tenido como telón de fondo la implementación computacional de estas cuestiones— son pocas las teorías que proponen una formalización viable en términos computacionales. La repercusión de estos fenómenos en la trama de la lengua ha de estar bien definida y precisada para la posterior automatización; la solución que se adopte para cada uno de estos temas será de aplicación homogénea a todos los procesos de etiquetación y deberá contemplar las limitaciones que un sistema computacional presenta en cuanto al modo de expresión, interpretación e integración de la información.

Dos de las fortalezas que el PLN puede ofrecer a la investigación lingüística teórica es, por un lado, el replanteamiento de viejas cuestiones desde nuevas perspectivas y, por otro, la aplicación de nuevas técnicas —basadas en la informática— para la descripción y análisis de la lengua. Este cambio de enfoque puede revelar, y de hecho revela, aspectos del hecho lingüístico que, cuando menos, obligan a la revisión de cuestiones que aún no han alcanzado una versión unánime y satisfactoria.

5.5.1.2 *Perspectivas gramaticales*

De acuerdo con Ducrot, una categoría lingüística, un paradigma, es generalmente mucho más que una colección o conjunto de elementos. Por lo común, comporta una organización interna e instituye entre sus elementos relaciones particulares (1974:136).

La amplitud de esta definición da lugar a la posibilidad de examinar esos elementos y relaciones desde distintas perspectivas. Por eso, cada escuela o

tradición elabora una taxonomía precisa de las categorías gramaticales que contarán en su concepción de la gramática. En su monografía sobre este tema, Bosque recoge cuatro CLASIFICACIONES BINARIAS de amplia aceptación en la tradición gramatical, que parten de criterios diferentes. Se oponen así, categorialmente, las palabras variables a las invariables, las que pertenecen a paradigmas abiertos a las que se deben a los cerrados, las palabras de las categorías llenas conceptualmente a las vacías, casi imposibles de definir, y las categorías mayores que pueden recibir complementos a las menores, que no presentan esa capacidad. (1990:28 y ss.).

Cabe una última clasificación binaria, desarrollada mayoritariamente en el marco de la gramática formal contemporánea y de la pragmática, que recoge algunas de estas generalizaciones e integra, además, otras categorías gramaticales menores como la flexión[54].

> […] el reconocimiento de las categorías funcionales presentes en la oración ha supuesto en los últimos años un considerable avance en los estudios que tratan los rasgos funcionales que relacionan los constituyentes oracionales (Ramalle, 2005:53).

5.5.1.3 Categorías léxicas y categorías funcionales

Esta clasificación se remonta, como tantas otras percepciones, a la *Poética* de Aristóteles, donde ya se advertía la existencia de *palabras con significado* y *palabras sin significado*.

Las CATEGORÍAS LÉXICAS, designadoras o denotativas, recogen bajo su epígrafe al *adjetivo, sustantivo, verbo* y *adverbio*. Remiten a realidades externas al sistema gramatical; designan cosas, personas, acciones, cualidades, modos de actuar, etc. Por este motivo son categorías abiertas, en constante recuento y, puesto que tienen significado léxico, pueden seleccionar a sus complementos. Son, en pocas palabras, las que manifiestan los significados que uno iría a buscar al diccionario.

[54] La gramática generativa actual ha sustituido el término tradicional de *categoría gramatical* por el de *funcional*. La oposición léxico-gramatical da a entender que las categorías léxicas están fuera del análisis de lo gramatical lo que es, a todas luces, incorrecto. No obstante, el término *funcional* no está libre de problemas debido a su polisemia en el ámbito de la lingüística estructuralista europea (Bosque y Gutiérrez, 2009:108 y ss.).
Véase Chomsky (1995) y Bosque y Gutiérrez (2009) para el enfoque de la gramática formal, y Sperber y Wilson (1993 y1994) y Escandell y Leonetti (2000), Escandell (2004) para la perspectiva pragmática.

Frente a estas, las CATEGORÍAS FUNCIONALES, representadas mediante marcas tales como la flexión verbal, la referencia, las conjunciones coordinantes y subordinantes, las preposiciones y los determinantes, constituyen clases cerradas que no pueden incrementarse. Una lengua puede tomar prestada de otra lengua una palabra de una categoría léxica, pero no aplicará este proceso de préstamo con una conjunción o un pronombre.

En general, los hablantes conocen y usan las categorías funcionales de su lengua de forma regular, mientras que el dinamismo natural de las categorías léxicas hace imposible que se conozcan y utilicen todas sus unidades. Los diccionarios, en este caso el lexicón, contienen mucha más información léxica de la que puedan conocer los hablantes, sin embargo, las categorías funcionales son esenciales en todo discurso, traspasando todas las variedades tanto geográficas como diacrónicas y registros lingüísticos que puedan presentarse en una lengua. Aunque no tienen significado léxico —se definen en función de los conceptos propios de la gramática—, pueden presentar rasgos funcionales como el género, el número, la persona y el tiempo que establecen vínculos mediante relaciones de *concordancia de rasgos* con las categorías léxicas.

Debido a su naturaleza procedimental transmiten información gramatical que contiene instrucciones que activan los procesos sintácticos contribuyendo significativamente a la interpretación de las categorías léxicas tanto en el plano semántico como en el pragmático. El significado de esta información nos indica cómo manipular las representaciones conceptuales de modo que se garantice cierta eficacia en el proceso de inferencia e interpretación[55]. Las categorías funcionales, por lo tanto,

> […] representan la capa o estrato funcional que se superpone a las proyecciones léxicas del verbo y del nombre, siendo las responsables de la organización estructural de los enunciados (Escandell y Leonetti, 2000:364).

[55] Su función es orientar hacia ciertas rutas que han de seguir los supuestos y las inferencias, al tiempo que vedan otros vericuetos no relevantes (o inadecuados) para la interpretación. El tipo de instrucciones que activan son de tipo computacional y afectan no sólo a la sintaxis sino, como se ha sugerido, a la interpretación (Escandell y Leonetti, 2000).

Por este motivo, no son independientes[56], no se presentan aisladas, son inherentemente transitivas, en un sentido amplio del término; necesitan que otras categorías —generalmente léxicas— llenen o saturen los espacios sintácticos que proporcionan (Bosque y Gutiérrez, 2009:112).

5.5.1.4 *Las categorías funcionales del etiquetario*

El trabajo que se ha llevado a cabo en este punto, parte de esta clasificación general de categorías léxicas y funcionales para hacerse cargo de las últimas que, a la postre, son los engranajes de las categorías *llenas, mayores, variables*, o de *paradigmas abiertos*. Estas unidades son las que mayor dificultad revisten para cualquier gramático o lexicógrafo, ya sea en la tarea teórica de descripción y análisis, o en la labor práctica de catalogación de cada unidad.

Las categorías funcionales que se considerarán serán las PREPOSICIONES, las CONJUNCIONES, los ARTÍCULOS, los PRONOMBRES y las unidades que en el discurso serán DETERMINANTES. Esta selección varía según se acepte la clasificación de uno u otro autor; de hecho este inventario *todavía es objeto de discusión y de análisis* (Bosque y Gutiérrez, 2009:114)[57].

Otro de los inconvenientes que se presentan en el proceso de formalización es que tanto los pronombres como los determinantes se nutren, en parte, de las llamadas CLASES TRANSVERSALES, que recogen ámbitos como el de los demostrativos, posesivos, cuantificadores, numerales, indefinidos, interrogativos, exclamativos y relativos. Esta característica dará lugar a clasificaciones cruzadas, por lo que se establecerá un conjunto de etiquetas de estas clases que sirva de base para el posterior etiquetado de cada unidad.

Teniendo en cuenta las limitaciones que constriñen el procesamiento automático de la lengua, se ha acudido a cuantas fuentes y marcos teóricos han sido necesarios para obtener una descripción y argumentación sólida que

[56] Exceptuamos de esta afirmación a algunos pronombres.

[57] Categorías de amplia aceptación en el marco generativista como la *flexión* o el *complementador* no serán consideradas, en este sentido, en el desarrollo de este trabajo. Estas categorías se activan en el eje sintagmático, en el plano del discurso, por lo que están fuera del ámbito de la etiquetación de un lexicón, donde el objetivo es la clasificación de las palabras en el paradigma.

permitiera una formalización e implementación satisfactoria en términos tanto lingüísticos como computacionales.

5.5.1.5 *Criterios para la identificación de unidades. Límites y fortalezas*

Siendo las categorías gramaticales las unidades básicas de la sintaxis, una teoría que pretenda dar cuenta de la estructura interna de la lengua no puede prescindir de una concepción clara y coherente de lo que son dichas categorías, cómo se articulan entre sí y cuáles son las diferencias y relaciones que existen entre ellas. Pero, como hemos intentado poner de manifiesto en la brevísima panorámica expuesta, las categorías gramaticales han sido pensadas y clasificadas desde muy distintos enfoques y criterios, aunque desde todos ellos laten las mismas preguntas; ¿qué número de categorías debe postularse?, ¿qué criterio debe primar en la delimitación?

> Casi todos los autores modernos reconocen que las listas de clases de palabras con las que trabajan están basadas en una extraña mezcla de criterios, […]. La paradoja habitual de las categorías gramaticales es precisamente que no existe autor ni escuela que no reconozca la dificultad de obtenerlas formalmente, mientras que a la vez son unidades básicas de análisis en casi todos los marcos teóricos (Bosque, 1990:23-24).

Todas las clasificaciones citadas muestran ciertas convergencias que sugieren una evidencia en cuanto al estado de esta cuestión concreta; las categorías gramaticales son susceptibles de ser identificadas, definidas y descritas a partir de un amplio manojo de criterios —morfológico, sintáctico, funcional, semántico, textual, etc.—, y perspectivas; estructuralismo, distribucionalismo, generativismo, funcionalismo, lingüística cognitiva, etc.

En nuestra gramática tradicional, han convivido estos criterios y perspectivas a la hora de la definición de nuestras categorías; mientras el verbo típicamente ha sido justificado desde un criterio puramente morfosintáctico, *aquellas palabras con conjugación que concuerdan con el sujeto en número y persona,* otras atendían a planteamientos claramente semánticos, *el sustantivo expresa individuos y objetos, y el adjetivo cualidades*. La distribución era necesaria para definir al pronombre, *palabra que sustituye al nombre,* y la función y semántica para dar cuenta del adverbio; *palabras que funcionan como complemento circunstancial que significan tiempo, lugar, modo*, etc.

Los criterios lingüísticos más utilizados para delimitar la nómina de categorías gramaticales han sido el DISTRIBUCIONAL, el FUNCIONAL, el MORFOLÓGICO y el SINTÁCTICO, aunque, como se ha visto, no son ajenas en nuestra tradición las definiciones NOCIONALES. La falibilidad de esta última pauta ha sido ampliamente demostrada; *invasión* es un sustantivo, pero denota acción, como los verbos, y *blancura* que es una cualidad no es un adjetivo. *Saber* o *tener* no son acciones, y sin embargo son verbos, y adjetivos como *este* o *aquel* no parecen denotar ninguna cualidad. No obstante, en ocasiones, será necesario acudir a consideraciones semánticas —*poder cuantificar, poder referir, tener capacidad mostrativa*— para caracterizar categorías como la de los determinantes. A pesar de lo limitado y problemático de este criterio, máxime teniendo en cuenta que la máquina no presenta la capacidad de interpretar significados, será necesario observarlo para determinar las clases transversales, así como algunas subclases presentes dentro de cada categoría.

Se presenta, a continuación, una brevísima nota sobre los criterios de identificación categorial más invocados en el ámbito gramatical y que han sido utilizados para llevar a cabo la catalogación y etiquetación de nuestro lexicón. Como se podrá advertir, se ha hecho necesario observar más de un criterio con el fin de obtener definiciones claras que delimiten con precisión las categorías y permitan ser reconocidas por la máquina.

Criterio distribucional

El criterio distribucional propone como axioma identificar las unidades a partir de su entorno. Algunos gramáticos aplicaron este método a la sintaxis y al estudio de sus unidades básicas, localizando los diversos contextos sintácticos en que puede aparecer una categoría. Concluyen así, que las palabras que tienen idéntica distribución y, por lo tanto, mismo entorno o cotexto, pertenecen a una misma clase.

Esta técnica, que alcanzó notables éxitos en los ámbitos de la fonología y la morfología, presenta ciertos inconvenientes en el nivel sintáctico. El principal es la confusión o indiferenciación entre categoría y función que puede desempeñar una categoría. Parece obvio que una misma función puede ser desarrollada por diferentes categorías sin que esto implique que esas categorías sean idénticas;

Nunca fuiste {sincero ~ alto}/{presidente ~ el presidente}/{quien dijiste ser}/{de cara ~ de confiar}/{demasiado listo}.

Los criterios puramente distribucionales, por lo tanto, solo pueden determinar qué categorías ocupan típicamente una determinada posición bajo determinadas circunstancias, pero este criterio no permite distinguir, por ejemplo, un pronombre de un nombre propio o de un infinitivo.

Criterio funcional

El funcionalismo tradicional, que se consolida en nuestra tradición hispánica gracias al paradigma alarquiano, ha caracterizado sus categorías a partir de las funciones que desempeñan. Las funciones, por lo tanto, son previas a la categoría.

Se trata de un análisis basado en la identificación entre la forma y la función para la caracterización de las categorías gramaticales. Este proceder elude, según Bosque, el problema de la clasificación al pasar por alto ciertas generalizaciones que emergen a nivel estrictamente categorial (1987). El criterio funcionalista propone, siguiendo esta lógica, que serán sustantivos no solo las formas que representan la categoría léxica de sustantivo, sino también las oraciones conmutables por un sustantivo, que son susceptibles de desarrollar las funciones propias de este; sujeto, complemento directo, complemento de régimen, etc. Pero hay evidencias que contradicen este *a priori*; existen verbos transitivos que admiten complementos directos nominales, pero no oraciones subordinadas sustantivas —que desempeñan esa misma función—; **Antonio pesa que venías otra vez*.

Criterio morfológico

Este análisis ha sido especialmente útil para discriminar las categorías que presentan marcas de flexión de aquellas que son invariables. De este criterio emerge la siguiente clasificación:

VARIABLES	flexión de número	SUSTANTIVO[58]
	flexión de género —masculino y femenino— y número;	ADJETIVO
	flexión de género —masculino, femenino y neutro— y número;	PRONOMBRE
	flexión de tiempo, modo, número y persona	VERBO
INVARIABLES	PREPOSICIÓN, CONJUNCIÓN, ADVERBIO, INTERJECCIÓN, CONECTOR DEL DISCURSO[59]	

Esta taxonomía, que resulta muy útil en su conjunto, no es exacta en sus detalles; hay adjetivos que no tienen morfema de género —*fácil*—, tampoco está presente en todos los pronombres —*yo*—, hay adverbios como *cerquita o prontito* que admiten flexión de grado, las palabras sin flexión están recogidas en una misma clase, etc.

Pongamos por caso la categoría adverbial. Como es bien sabido, dentro de esta clase se incluyen palabras de naturaleza y distribución bien diferentes, que aun pudiendo coaparecer en un mismo entorno, no son intercambiables.

Llegó muy rápidamente ayer.
Llegó rápidamente ayer.
Llegó muy rápidamente.
Llegó rápidamente.
Llegó ayer.
**Llegó muy.*
**Llegó muy ayer.*

Este breve ejemplario pone de manifiesto lo diferente que es la gramática de unidades adverbiales como *rápidamente* o *muy,* formas que parecen compartir exclusivamente la ausencia de flexión.

Por lo expuesto podemos advertir que ni el criterio basado en la distribución, ni la identificación entre función y forma, ni el criterio morfológico resuelven por sí

[58] El sustantivo tiene género, pero no variación o morfema flexivo. Aun así rige, gobierna o selecciona un determinado morfema de género con el que concordar.

[59] En opinión de algunos autores, los marcadores del discurso deben incluirse en este grupo de categorías. (Martín Zorraquino y Portolés 1999, GDLE § 63.1; Pavón, 2003).

solos el problema ilustrado; la ausencia de una taxonomía consensuada y precisa que sirva de clasificación para las unidades que encarnan la oración.

Criterio sintáctico

Para la identificación rigurosa de las categorías necesitamos el apoyo de, entre otros, criterios sintácticos basados en el estudio de las propiedades sintácticas que caracterizan a las diversas unidades agrupadas bajo una misma etiqueta categorial. En la actualidad, habida cuenta de las limitaciones que presentan los enfoques anteriores, se tiende a definir las categorías gramaticales a partir de su comportamiento dentro de la oración, esto es, según las relaciones que establezcan sintagmáticamente con las demás categorías; si llevan o no modificaciones y complementos, si estos están exigidos o no por el significado de la pieza léxica, etc. (Ramalle, 2005:38).

Una catalogación basada en criterios sintácticos surgirá, por lo tanto, del ANÁLISIS DE LAS RELACIONES que establece una pieza léxica con el resto de categorías con las que coaparece y de las MARCAS FORMALES que estas relaciones dejan impresas en el desarrollo de la oración.

5.5.1.6 Flexión y sintaxis

Ch. Hockett en su *Curso de lingüística moderna* definía las categorías gramaticales enlazando estas dos dimensiones:

> Las partes de la oración se definen como clases formales de temas contenientes que muestran comportamiento similar en la flexión, la sintaxis o ambas (1971:225).

Como sostiene Varela (1999), morfología y sintaxis tienen un vocabulario compartido. Son dos aspectos que se presentan en todas las estructuras, tanto a nivel oracional, como en el de la palabra. Las relaciones entre ambos niveles se pueden resumir en dos líneas básicas de análisis.

Por un lado, ciertos procesos morfológicos tienen que hacer uso de principios sintácticos para dar cuenta, por ejemplo, de la selección que un determinado sufijo realiza de su base derivativa. En el interior de la palabra se advierten mecanismos sintácticos que se manifiestan en procedimientos como la composición o derivación léxica. Según Varela, los sufijos no solo poseen información categorial;

también pueden contener otros rasgos morfosintácticos relevantes tales como información temática sobre la base a la que se unen. Un ejemplo de restricción morfosintáctica se encuentra el sufijo –*dor,* que solo está presente en los nombres derivados de verbos que tienen argumento externo; *fumar→ fumador, vencer→ vencedor.* Son incompatibles, por lo tanto, con verbos inacusativos, es decir, con argumento interno; *terminar→*terminador, existir→*existidor,* etc. (Varela 1999:274 y ss.).

Por otro lado, la sintaxis también tiene que tener en cuenta aspectos relacionados con la morfología. Un ejemplo de esta interdependencia lo constituyen los nombres de acción que proceden morfológicamente de verbos y se construyen como sus bases derivativas con complementos que representan el agente y el paciente.

Juan (agente) solucionó el problema (paciente).
La solución del problema (paciente) por parte de Juan (agente).

Tanto *solucionar* como el sustantivo derivado *solución,* comparten un mismo esquema sintáctico de dependencias. Y esto es así porque los sustantivos deverbales heredan del verbo del que proceden su estructura argumental (Ramalle, 2005:145 y ss.).

Como se observa, ambas dimensiones, flexión y sintaxis, sirven para identificar a las categorías gramaticales que se reconocerán tanto por su función en la oración, como por sus marcas formales. Las relaciones sintácticas entre los elementos suelen dejar marcas en la palabra a través de los morfemas flexivos, que en este sentido son marcas de función. Acéptese como ejemplo la palabra *prueba,* que no muestra su categoría gramatical si no es incluida en su contexto. Su flexión aporta información sobre sus potencialidades —sustantivo y verbo— pero no es suficiente. El criterio morfológico no desambigua las formas; es la sintaxis la que ancla las palabras a su categoría final.

Entendemos pues, con Varela, que para establecer un repertorio de categorías son imprescindibles ambos análisis, tanto el ofrecido por la morfología, crucial para elucidar la categoría gramatical, como el sintáctico, que precipita el desenlace atribuyéndole a cada forma su función.

Estos criterios, por lo tanto, han sido los ejes que han motivado el diseño del etiquetario que se ha seguido para la clasificación y anotación manual de las categorías funcionales del Lexicón TIP.

Una reseña aparte merece la categoría funcional de los determinantes que, como se podrá colegir a partir de las páginas de este trabajo, se ha revelado como un instrumento esencial para el proceso de formalización y codificación de las estructuras que dan contenido a los patrones codificados de error. Por otro lado, los determinantes se ofrecen como un instrumento esencial para la desambiguación de las formas. Será necesario, por lo tanto, afrontar un estudio exhaustivo de corte teórico que permita acotar el territorio propio de los determinantes con el fin de esclarecer y precisar tanto sus límites, como sus unidades. Este estudio deviene en el desarrollo de un inventario cerrado[60] que recoge todas las formas constituyen y cierran cada paradigma, y puede servir de herramienta o guion para la anotación de otros lexicones en nuestra lengua.

5.6 Los determinantes

La categoría funcional de los determinantes es quizá uno de los capítulos más complejos y escurridizos de la gramática descriptiva del español. A pesar de que se ha escrito mucho sobre la naturaleza, función y elementos que integran el paradigma, la noción de determinante, comparada con otras como las de verbo, régimen o concordancia es relativamente joven. De hecho, no aparece en la gramática tradicional hasta los métodos estructuralistas y formales. Hoy, sin embargo, es un concepto básico de uso corriente y aceptación generalizada, tanto en el seno de la teoría sintáctica, como en el campo de la gramática descriptiva.

Su éxito, como afirma Leonetti, se debe a que es un concepto muy útil que permite simplificar enormemente la descripción gramatical (2000:11). Esta propiedad, tan deseable en cualquier paradigma o método de análisis, se hace esencial en la elaboración de reglas formalizadas. Y es por este motivo, simplificación en la

[60] Aunque ciertamente es un inventario cerrado, la presencia en este ámbito de la determinación de los adjetivos determinativos numerales hace, que al menos en ese paradigma, la puerta de entrada siempre quede abierta.

descripción gramatical que agiliza notoriamente el proceso de codificación, por el que se incluye esta categoría funcional en nuestro etiquetario, a pesar de no estar recogida como tal en las fuentes lexicográficas de referencia.

Los determinantes constituyen un constructo necesario para ciertas fases y procedimientos del PLN. Son una herramienta valiosísima para la tarea de desambiguación. De hecho, el bloque más amplio de ambigüedad en español, «N−Vb no finito», está constituido por palabras como *juego, venda, menta o meta*, que podrán desambiguarse en su gran mayoría a partir de la elaboración de reglas que contemplen la presencia o ausencia de un determinante; {*El/Mi/Algún*} *juego*; {*Yo/Cuando*} *juego*[61]. Por otro lado, y siguiendo las recomendaciones de los estándares de catalogación y etiquetación de EAGLES, será una categoría computable en las fases de POS *tagging* y análisis sintáctico.

5.6.1 Consideraciones previas

Como se justificó en la propuesta de etiquetario, consideraremos el grupo de determinantes no tanto como una categoría en el sentido en el que el adverbio, el verbo o el sustantivo lo son, sino como una clase de categorías que incorpora unidades de diferente naturaleza —artículos y adjetivos determinativos— que presentan morfologías y semánticas distintas. Estas formas se aliarán por tener una distribución y función sintáctica común, y por operar en un mismo nivel lingüístico, el discursivo. La función que comparten los determinantes es la capacidad de saturar la proyección nominal que habilita al sustantivo como argumento del verbo y la distribución, el entorno que ocupan, será el margen prenominal.

Por este motivo, la etiqueta de determinante no formará parte del repertorio manejado para el lexicón, aunque esta información funcional será de gran utilidad para el desarrollo del repositorio de errores; 301 patrones codificados en la base

[61] No obstante, habrá casos como el de la frase *la venda* en los que el determinante no sea útil para la desambiguación debido a que él mismo también es ambiguo con respecto al pronombre de complemento directo *la*. Las dos posibles secuencias que genera la combinación de estas unidades son legítimas; «ART + N» y «PN + Vb», por lo que no será posible aplicar a estos casos la regla general; *si ante la ambigüedad* «N−Vb no finito» *hay una palabra en función de determinante, la solución a la ambigüedad será N.*

de datos de errores tienen presente esta categoría en su entrada. Esta categoría, como se verá, facilita por un lado, la desambiguación y por otro, la identificación del error.

A continuación, se expondrán las bases para la construcción de una clasificación que recoja esta categoría discursiva en el marco del PLN. Se hará una descripción de los determinantes en español orientada a la elaboración de una taxonomía que permita la codificación y etiquetación de todas estas formas. Este trasfondo computacional no obstará para que la clasificación se ajuste satisfactoriamente a los rigores de la teoría gramatical.

5.6.2 Las formas potenciales: problemas relacionados

Para el abordaje del repertorio de determinantes del español, las gramáticas suelen partir de la aceptación de que los artículos, los demostrativos y los posesivos son elementos constitutivos de esta clase gramatical. Si se interpreta el término en un sentido amplio, de forma que dé cabida a otras unidades que legitiman a los sustantivos como argumentos de un predicado, la clase de los determinantes acoge también a cuantificadores prenominales como *alguna, tres, muchos*, *etc*. (NGLE §1.9r).

Esta segunda interpretación, más abarcadora, será la asumida en esta propuesta, aunque debe hacerse una salvedad; no consideramos aquí que demostrativos, posesivos o cuantificadores sean categorías gramaticales, sino más bien membretes que responden a criterios semánticos. La información que revelan de la palabra no es de tipo categorial, sino semántico y funcional. Entendemos por lo tanto, que las palabras adscritas a estos grupos, que llamaremos como en la NGLE *clases transversales,* necesitan ser etiquetadas, en primera instancia, con una categoría gramatical, que permitirá su tratamiento por parte de la máquina en las mismas condiciones que el resto de las palabras que forman parte del lexicón[62].

[62] En los estudios sintácticos actuales, tanto teóricos como computacionales, suele considerarse que todas estas informaciones gramaticales (nombre, cuantificador, demostrativo, etc.) son rasgos gramaticales de la palabra. Se parte de la idea de que las piezas léxicas son una matriz de rasgos de distinta naturaleza; categoriales, funcionales, de subcategorización, etc. Uno de los problemas para construir un *sistema de haces de rasgos* es elegir cuáles son los elementos fijos del paradigma y cuáles los transversales.

Los *adjetivos*, siempre que respeten el margen prenominal y los criterios sintácticos que se explicitarán líneas más abajo, podrán desarrollar, junto con los *artículos*, la categoría de determinante. De acuerdo con los criterios expuestos por Eguren y Soriano, se hará diferencia entre la función de los DETERMINANTES y la etiqueta de DETERMINATIVO. Este rasgo semántico lo desarrollarán los adjetivos que están vinculados a las operaciones de cuantificación y referencia de un elemento nominal (2006:21 y ss.).

5.6.3 Los adjetivos determinativos

El término adjetivo se suele emplear en un sentido laxo y en otro restrictivo. El primero, más frecuente en los estudios tradicionales, es el resultado de privilegiar los dos criterios formales que caracterizan a esta clase de palabras: la *concordancia* con el sustantivo y su función como *modificador* de este. Considerado este sentido abarcador del adjetivo, se da cabida en esta clase de palabras a todos los elementos subrayados en la relación siguiente: *esas plantas carnívoras; muchos buenos amigos; cuya absurda opinión; un día cualquiera.*

Nuestra gramática tradicional ha dividido los adjetivos marcados en CALIFICATIVOS y DETERMINATIVOS. Los primeros, —*carnívoras, buenos, absurda*— designan cualidades y representan mayor libertad en su distribución pudiendo aparecer en ambos márgenes del sustantivo. Los determinativos, por su parte, son los que delimitan la denotación de un grupo nominal especificando a cuántas y cuáles de las entidades designadas por el nombre hace referencia el hablante. Bajo esta etiqueta se suelen subsumir formas de las clases transversales como la de los demostrativos, los posesivos, los indefinidos, los numerales, etc. Estos adjetivos pueden ocupar tres posiciones dentro del grupo nominal;

Su voz.
La voz suya.
La otra voz.

Al margen de estos debates, la pauta que se seguirá en la etiquetación del lexicón será aportar el máximo de información —ya sea gramatical, funcional o semántica— de modo que la descripción de cada ítem tratado computacionalmente sea lo más rica, relevante y universal posible.

pero solo las palabras situadas en posición prenominal podrán ejercer la función de determinante[63]; las pospuestas actuarán como modificador del sustantivo.

Este cisma en la categoría de los adjetivos —calificativos y determinativos— es, a pesar de su tradición, una división controvertida debido a las evidentes diferencias entre estos dos subgrupos. Como señala Leonetti;

> Los adjetivos calificativos se combinan con los nombres para permitir la construcción de nuevas subclases de objetos, y los determinativos lo hacen para permitir la extracción de determinados objetos o cantidades de objetos dentro de la clase denotada. La contribución de ambos tipos de elementos a la interpretación de los constituyentes que los contienen no es, por tanto, la misma (2000:14).

Estas evidencias, no obstante, no han sido suficientes para que muchas de las gramáticas consultadas, tanto descriptivas como computacionales, así como las fuentes lexicográficas recogidas en la base de datos, hayan llevado a cabo de forma más drástica esta distinción. Unas y otras formas, a pesar de sus diferencias tanto semánticas, como morfosintácticas, siguen cayendo en la misma clase; los adjetivos.

Aunque este tipo de debates y revisiones corresponden más a la lingüística teórica que a la computacional, creemos necesario evaluar estos datos. No en balde, y debido a las controversias suscitadas por esta subclase, la propuesta de EAGLES que parte de una subclasificación de los adjetivos como la que se viene exponiendo, finaliza con la siguiente recomendación:

> Each language-specific application should specify clearly in which category the indicative adjectives are treated in order for cross linguistic comparisons to be made possible (1996, *Type* "Preliminary Recommendations").

La utilidad de la catalogación y etiquetado tanto de un lexicón como de un corpus, se mide en buena medida, según el grado de legibilidad y proximidad que presente con respecto a la teoría lingüística general. Pongamos por caso una aplicación de traducción automática; aunque en términos estrictamente

[63] La especialización que presentan estos adjetivos en la función de determinar (frente a la típicamente adjetiva de calificar) ha producido que muchas de estas voces hayan sufrido o estén en ciernes de un proceso de gramaticalización que conlleva la pérdida de parte de su sentido adjetivo original en favor de propiedades características de los determinantes. Adjetivos como *semejante, propio* y *cierto* que son considerados determinativos, revelan en su uso actual este proceso de gramaticalización.

gramaticales parece evidente que la tradicional categoría de los adjetivos recoge formas que pertenecen a clases diferentes, deben observarse otros criterios como el tipo de etiquetas que mayoritariamente se empleen para el tratamiento de la otra lengua, intentando salvaguardar, en la medida de lo posible, las mismas bases de catalogación de modo que se facilite la transferencia de datos.

5.6.4 Adjetivos y pronombres: la duplicación de paradigmas

Junto a las oposiciones citadas que distancian a las dos subclases de adjetivos, cabe un último matiz diferenciador; las formas de los determinativos son susceptibles de un uso pronominal, esto es, de duplicar su paradigma para encarnar una nueva función. Pueden, por lo tanto, aparecer como expresiones independientes que no determinan ni modifican a ningún nombre;

Alguien llamó a la puerta.
Vino con todo quemado.
No entiendo que otro haya podido ocupar mi lugar.
Se juntó con unos peores que él y arruinó su vida.

En estos casos, muchos análisis gramaticales tratan las palabras marcadas como *determinantes*, *adjetivos con uso pronominal* o *formas adjetivas de los pronombres,* entre otras. Otra solución, a la que se suma la NGLE[64], es la propuesta por las tendencias sintacticistas, que plantean un análisis alternativo que evita duplicar el paradigma de los determinativos[65]. Estos autores consideran que en estos usos pronominales hay un elemento tácito cuyo contenido se recupera del contexto —un sustantivo o grupo nominal— al que se refiere el adjetivo determinativo; *Vino con todo Ø quemado; Se juntó con unos Ø peores que él y arruinó su vida.*

En el sistema que aquí se plantea, estas formas responderán a la etiqueta de pronombre por lo que será necesario doblar estos paradigmas. La duplicación es un recurso tradicional en el análisis sintáctico que está habitualmente asociado a

[64] Según la RAE los dos criterios, tanto el sostenido por la tradición (multiplicación de paradigmas) como el que apuesta por la existencia de un elemento tácito, son válidos y proporcionan análisis equivalentes, pero se prefiere el segundo por cuestiones de simplificación y elegancia descriptiva (NGLE §17.2g).

[65] Lázaro (1980), Hernanz y Brucart (1987), Garrido (1996).

las unidades que poseen capacidad referencial. Aunque hay alternativas solventes como esta última de la elisión, se recurrirá aquí al recurso de la duplicación, siguiendo la recomendación de EAGLES de diferenciar la categoría de los pronombres de la de los determinantes. Esta escisión, que puede plantear debate en el plano teórico, está justificada en el ámbito computacional, porque como advierte Civit:

> Desde la perspectiva de PLN es conveniente distinguir determinantes de pronombres en los primeros niveles de análisis. Las estructuras en las que intervienen ambas clases de palabras son distintas, y si la correcta etiquetación puede hacerse en los niveles morfológicos en vez de esperar a los sintácticos, mejor, porque se reduce la ambigüedad en los niveles de análisis posteriores (2003: 85).

Por este motivo consideramos que tanto *posesivos* como *demostrativos, cuantificadores, indefinidos, numerales, relativos, interrogativos* y *exclamativos* constituyen clases que, desde el punto de vista funcional y morfosintáctico, son encuadrables tanto en la categoría de pronombre, como en la de adjetivo determinativo. Evitamos así, en la medida de lo posible, la interpretación de elementos nulos por parte de la máquina.

Esta forma de proceder —duplicación de paradigma—, como advierten muchos autores, introduce un notable grado de redundancia en el sistema gramatical e impide el descubrimiento de generalizaciones que podrían contribuir a una descripción más simple y elegante de estas palabras[66]. No obstante, es como se viene manteniendo, la solución más idónea en términos computacionales, porque además de facilitar la tarea de análisis, no rebaja la eficacia de la máquina a la hora de procesar las unidades de la lengua.

5.6.5 Gramática de los determinantes

En cuanto al régimen sintáctico de los determinantes, como categorías funcionales que son activan procesos sintácticos que contribuyen significativamente a la interpretación de las categorías léxicas. El determinante, para activar estos procesos, concuerda en género y número con el núcleo nominal al que afecta, y es

[66] Leonetti elabora y justifica la idea intuitiva de que los determinativos constituyen una única clase de palabras, independientemente de que funcionen en unos contextos como *adjetivos* y en otros como *pronombres* (2000).

indicador de estas marcas en aquellos casos en los que el sustantivo es invariable*; {el/los} lunes, {el/la} protagonista*[67].

Los determinantes serán considerados, por lo tanto, como piezas léxicas que subcategorizan una proyección con función nominal y deciden su referencia indicando qué entidades de las pertenecientes al conjunto denotado por la proyección del nombre deben tomarse en consideración al interpretar la secuencia. Su acción, pues, se extiende a varios niveles; selecciona el componente léxico, habilita al sustantivo a nivel sintáctico y orienta la interpretación en lo semántico.

En términos de distribución en el eje sintagmático, los determinantes son, en general, constituyentes obligatorios del grupo nominal y son pocos los contextos en que es posible su omisión. Su posición, sujeta a variación paramétrica, ocupa en español el margen izquierdo del sustantivo, en la posición más exterior, de modo que pueda individualizar y otorgar referencia a todo el conjunto*.*

A diferencia de otros modificadores del sustantivo, que pueden concurrir ilimitadamente bajo una disposición jerárquica que respeta el *criterio de la pesantez*[68]*,* las posibilidades de aparición de varias unidades en la posición del determinante queda restringida a casos muy concretos. No todos pueden aparecer en cualquier contexto; existen dependencias y restricciones de coaparición, así como otras restricciones seleccionales impuestas por efectos universales como el de *definitud* o *especificidad*.

Podemos vislumbrar, no obstante, ciertas generalidades; los determinantes definidos presentan distribución complementaria. Así, existe incompatibilidad entre el artículo, el demostrativo y el posesivo en la posición prenominal de determinante **el este muchacho; *la tu sonrisa*, no así en posición postnominal, en la que el adjetivo determinativo funcionará como modificador: *el chico ese.* En cuanto a los cuantificadores, pueden concurrir con los demás tipos de

[67] Esta, entre otras características, conduce a muchos autores de corte generativista a considerar que el determinante es el núcleo de una categoría funcional SDET, encargada de aportar al nombre la capacidad de referencia a través de mecanismos como la *determinación* o la *cuantificación*. Puede verse en Chomsky (1982), Abney (1987), Eguren (1989), Escandell y Leonetti (1997), Ramalle (2005), Bosque y Gutiérrez (2009).

[68] Hernanz y Brucart (1987:168).

determinantes, aunque también existen restricciones de aparición basadas tanto en criterios sintácticos como semánticos.

Procediendo a grandes trazos, podemos distinguir dos grandes grupos de determinantes según su contexto posible[69]:

- Aquellos que ocupan obligatoriamente la primera posición. Estos son los artículos, los posesivos, los demostrativos y algunos cuantificadores como *algún, ningún, ambos, cualquier*, etc. Algunas de estas formas pueden relegarse a una segunda posición al ir precedidas del predeterminante universal *todo*[70].
- Los que pueden ocupar indistintamente la primera posición o cualquiera de las interiores. Este grupo lo integran los numerales cardinales y algunos cuantificadores e indefinidos como *otro, mucho, poco,* etc.

5.6.6 Los elementos del paradigma

El elenco de determinantes constituye un paradigma cerrado de palabras que no puede incrementarse mediante creaciones léxicas nuevas[71]. Es posible, por lo tanto, ceñir el campo léxico de las unidades que podrán desarrollar esta categoría y ofrecer un listado que contenga todas estas formas. Para ello, partiremos de CRITERIOS FUNCIONALES —habilitar a un grupo nominal como argumento dentro de una oración—, DISTRIBUCIONALES —inscribirse en el margen prenominal— y SINTÁCTICOS.

Según Eguren (1989) hay dos pruebas sintácticas que diagnostican qué palabras pueden ser consideradas determinantes en español.

Por un lado, son elementos necesarios para fijar la referencia del N; su ausencia en ciertas posiciones causa agramaticalidad; **Lectura de este libro es muy recomendable*. Por otro lado, pueden aparecer solos encabezando el grupo nominal

[69] Hernanz y Brucart (1987:185 y ss.).

[70] Se trata de un cuantificador universal de naturaleza especial que incluye al resto de unidades que forman parte del grupo nominal, incluido el artículo. Quizá por este motivo se sitúe en la posición más a la izquierda del SN.

[71] De esta afirmación, como puede intuirse, debe excluirse el ámbito cuantificador de los numerales.

sin necesidad de apoyarse en otros determinantes. Siguiendo estas pautas de Eguren, las unidades que cumplen con ambos criterios son los ARTÍCULOS, los ADJETIVOS DEMOSTRATIVOS, los POSESIVOS, los RELATIVOS, INTERROGATIVOS y EXCLAMATIVOS, los NUMERALES CARDINALES, algunos CUANTIFICADORES y algunos INDEFINIDOS.

No se consideran determinantes, en cambio, adjetivos como *mismo, anterior, propio, enésimo,* etc., ni los numerales ordinales, porque necesitan el apoyo de otro determinante para formar secuencias gramaticales; **Propia ~ Misma Conchita no se lo creía.*

Establecidas estas bases, se ha llevado a cabo un análisis minucioso de cada forma junto con sus posibilidades combinatorias y su distribución. Se ha tenido especial cuidado en aquellos paradigmas que presentan series átonas o formas apocopadas para el masculino, porque habrá casos en los que ciertas formas, siendo adjetivos determinativos no podrán ser determinantes. Así, se recogerán en este grupo formas como *algún, mi*, y *veintiún*, pero no otras como *alguno, mío o veintiuno*, debido a que estas últimas tienen vetado el margen prenominal.

La tabla de determinantes que proponemos bajo estas consideraciones es la siguiente:

<table>
<tr><td rowspan="2">ARTÍCULOS</td><td>DEFINIDOS</td><td></td><td></td></tr>
<tr><td>INDEFINIDOS</td><td></td><td></td></tr>
<tr><td rowspan="10">ADJETIVOS DETERMINATIVOS</td><td>DEMOSTRATIVOS</td><td></td><td></td></tr>
<tr><td>INDEFINIDOS</td><td></td><td></td></tr>
<tr><td>RELATIVOS</td><td></td><td></td></tr>
<tr><td>INTERROGATIVOS</td><td></td><td></td></tr>
<tr><td>EXCLAMATIVOS</td><td></td><td></td></tr>
<tr><td>POSESIVOS</td><td>Serie átona + nuestro y vuestro</td><td></td></tr>
<tr><td rowspan="3">CUANTIFICADORES</td><td>TOTALES</td><td></td></tr>
<tr><td>PARCIALES</td><td>Versión apocopada</td></tr>
<tr><td>NULOS</td><td>Versión apocopada</td></tr>
<tr><td>NUMERALES</td><td>CARDINALES</td><td>Versión apocopada</td></tr>
</table>

5.7 Las clases transversales

Las clases transversales están nutridas por palabras de distinta categoría gramatical pero mismo significante, lo que implica que la forma canónica en la mayoría de estos casos tendrá presencia en varios paradigmas.

Las categorías gramaticales de nuestro etiquetario que podrán formar parte de estas clases son los ADJETIVOS, PRONOMBRES, ADVERBIOS Y SUSTANTIVOS —estos últimos, solo en el caso de algunos numerales—.

Es histórica, en los estudios gramaticales de nuestra lengua, la controversia generada en cuanto a la clasificación categorial de los elementos que configuran las clases transversales; la pluralidad de propuestas de análisis así lo confirma. Se adopta aquí el concepto y campo de aplicación, aunque con matices justificados en el esquema de anotación del lexicón, de la NGLE (§1.9l) para recoger ámbitos como el de los DEMOSTRATIVOS, POSESIVOS, CUANTIFICADORES, INDEFINIDOS, NUMERALES, RELATIVOS, INTERROGATIVOS y EXCLAMATIVOS.

De acuerdo con las premisas que se han expuesto, y siguiendo la tendencia de la tradición gramatical en lo que a duplicación de paradigmas se refiere, entendemos que hay al menos dos categorías léxicas —que se excluyen recíprocamente— que desarrollarán la función demostrativa, la posesiva, la cuantificadora, la de los indefinidos, la de los interrogativos y exclamativos, la de los numerales y la de los relativos; estos son, los adjetivos —determinativos— y los pronombres. Así, en oraciones como *Buscaba aquella camisa de flores y al final encontró esta,* la forma demostrativa *aquella* será un adjetivo determinativo, mientras que *esta* será un pronombre.

En algunos casos será necesario triplicar una misma forma canónica para garantizar la coherencia de todo el sistema —*todo*—. La ambigüedad formal será pues, una constante en la clasificación y etiquetación de estas clases transversales. No obstante, consideramos que es esencial sacrificar esa simplicidad descriptiva de la gramática teórica —deseable en toda teoría científica— en favor de la precisión y la cobertura del lexicón. Por lo tanto, creemos como propone EAGLES, que una gramática computacional será más robusta y simple a la larga, si se acepta que en el sistema gramatical son necesarios tanto pronombres, como adjetivos.

En el ámbito sintáctico, entendemos que estas formas acogidas en las clases transversales pueden, bien combinarse con un grupo nominal, bien aparecer solas. En el primer caso, caben dos posibilidades:

- El adjetivo está pospuesto y funciona como un modificador del sustantivo; *El chico este.*
- El adjetivo desarrolla la función de determinante en posición prenominal; *Aquel triste invierno.*

En el segundo caso, el elemento determinativo será un pronombre que no requiere —aunque acepta— complemento; *Alguien encontrarás; Este de mi derecha es el que más me gusta.*

A partir de estos fundamentos y del examen minucioso de cada forma se ha desarrollado este repertorio de clases transversales dándole contenido a cada elenco con las formas que pueden desarrollar estas etiquetas semánticas. Este repertorio se ha organizado por categorías gramaticales de modo que puedan extraerse con precisión todos los elementos que forman parte de los determinantes, de los adjetivos determinativos y de los pronombres, categorías funcionales que ha resultado ser, como se ha discutido en estas páginas, las más complejas de ceñir y etiquetar. El contenido concreto de cada paradigma se ofrecerá en los siguientes Anexos:

(1) Cuantificadores

(2) Numerales

(3) Indefinidos

(4) Posesivos

(5) Relativos

(6) Interrogativos y exclamativos

(7) Demostrativos.

Capítulo 6

PatErr: estrategias para su constitución

El área dedicada al procesamiento automático textual acoge a un amplio elenco de aplicaciones dedicadas a la elaboración, gestión y revisión documental. Dentro de estas posibilidades, una de las aplicaciones más extendidas y que cuenta con un mayor número de usuarios dentro de las tecnologías lingüísticas aplicadas a la lengua escrita son los programas de verificación y corrección ortográfica, gramatical y de estilo. Podemos describir estos programas genéricamente como *herramientas de ayuda a la escritura* que tienen como objeto la mejora en la calidad de los textos producidos a partir del ordenador[72].

6.1 Planteamientos generales

A partir de los trazos que se ofrecieron en la Introducción puede intuirse que la idea que subyace de esta propuesta es, asentados en el enfoque que hemos denominado *lingüística de errores con fines computacionales*, desarrollar un repositorio de patrones de error codificados —PatErr— que sirva como recurso para programas o aplicaciones de revisión textual. Su basamento, como se ha ido anunciando, es la identificación de errores en el texto a partir del contraste con la información recogida en un repositorio que albergue patrones de error.

Para superar ciertas inconsistencias que se han detectado en otros programas similares, —recuérdese el caso ilustrado de «*a* + participio»—, los patrones de error que aquí se plantean deben ser capaces de abordar, en un contexto de contigüidad, no solo los casos erróneos concretos, sino generalizaciones y abstracciones de estructuras erróneas que encapsulen idealmente toda la casuística posible relacionada con el fenómeno al que se le pretenda dar tratamiento.

[72] Gómez Guinovart, (1999, 2001); Llisterri, (2003).

A partir de estos planteamientos, será posible ofrecer cobertura a todas las posibilidades que puedan derivar de un mismo fenómeno, con independencia de la frecuencia de aparición en la lengua de las palabras que involucran el error. Así, un error habitual como el citado «*a + participio*» será reconocido en PatErr con independencia de la frecuencia que presente el participio que forma parte de la estructura. De este modo serán tratadas secuencias como **a visto, *a tenido, *a venido*, del mismo modo en que serán reconocidos los errores en **a reverdecido* o **a ademado*.

El depósito de patrones de error, pues, actuará como el espejo en el que se refleja el texto que pretenda revisarse. Una vez detectado e identificado el error, el sistema deberá tener prevista una batería de acciones correctivas automatizadas que, básicamente, se concretan en la extracción de la cadena errónea y la reposición de la secuencia en su forma idónea ajustándose a lo normativo y, cuando sea preciso, a la propia naturaleza del texto.

Llevadas a cabo las medidas de subsanación, el sistema que se nutra de PatErr podrá ofrecer al usuario una glosa en la que se argumente la corrección y se aporten, además, ejemplos, generalizaciones gramaticales y referencias normativas que justifican el tratamiento que se ha llevado a cabo. El ofrecimiento de estas glosas —servicio de asesoría— supone una novedad con respecto a otros programas de revisión textual.

Este valor añadido podrá ser explotado en términos didácticos o pedagógicos. PatErr podrá dar soporte a una aplicación de verificación de errores que funcione como un manual de uso de la lengua capaz de abordar los casos en los que se ha constatado, a partir de la bibliografía especializada y de los datos arrojados por el Corpus TIP, que entrañan dificultad para los usuarios del español.

Aunque este planteamiento resulte simple en comparación con el desarrollo de una gramática computacional cuajada de reglas y restricciones capaz de procesar tanto lo correcto como lo erróneo, el reto sigue siendo titánico; crear un arsenal de patrones de error codificados que sirva como recurso eficiente y fiable para un sistema de identificación de errores del español escrito.

Conviene insistir, en este punto, que PatErr recoge y ofrece tratamiento a errores que se producen en un entorno de contigüidad. Los procedimientos de los

que hace uso se inscriben dentro de las técnicas que hemos descrito de *bajo nivel* por lo que no se llevarán a cabo análisis más abstractos que eluciden la estructura de los constituyentes de una oración y permitan pensar en tratamientos que traspasen la barrera de lo inmediato.

No obstante, como se intentó poner de manifiesto en el Capítulo 3, los sistemas de PLN más robustos son precisamente aquellos que incorporan únicamente conocimiento de los niveles más bajos de la descripción lingüística —morfología y sintaxis de bajo nivel—. Por lo tanto, cuanto mayor sea la profundidad del análisis del programa, que incorporará información semántica y pragmática, menor será su fiabilidad y su margen de manipulación de los datos. Es por esto que, según Ramírez Bustamante *et al. para el diseño de un corrector robusto debe considerarse únicamente información morfosintáctica y, en menor medida, semántica* (1994:575).

6.2 Concreciones

Para que PatErr, asociado al sistema de interpretación de patrones que se tiene previsto, pueda ofrecer ciertas garantías en términos de cobertura será necesario que esté lo suficientemente nutrido como para representar e identificar las incidencias que con mayor frecuencia se registran en el español escrito[73].

Por otro lado, para asegurar esa cobertura total, *consistente*, sobre los casos que derivan de un mismo fenómeno que ha sido seleccionado para su tratamiento será necesario, a partir de las descripciones gramaticales y de los resultados que arroje el Corpus TIP, un estudio exhaustivo del fenómeno seleccionado que permitirá ceñir los entornos y contextos de todos los casos que puedan asociarse a la misma incidencia.

Una vez acotado el fenómeno de error y listadas sus posibles derivaciones, se presentan dos opciones para llevar a cabo la tarea de elaboración de patrones.

[73] En el caso del trabajo que estamos presentando que, finalmente está pensado y desarrollado con el objetivo de integrarse competitivamente en el mercado de las industrias de la lengua, se hace necesario garantizar un amplio grado de cobertura y precisión en un tiempo de ejecución computacional razonable.

En primer lugar, se puede optar por incluir manualmente en una base de datos todos los errores que se detecten asociados al fenómeno. En un caso como el ya expuesto,

Patrón del error	Patrón de la corrección
a + participio	**ha** + participio

sería necesario elaborar tantos patrones como verbos existan, de modo que quede recogida toda la casuística que puede presentarse replicando este error.

La otra opción que se nos ofrece será elaborar patrones codificados que permitan bien generar, de alguna forma automática, todos estos errores posibles —junto con su corrección—, bien crear estructuras abstractas figurativas que sean interpretables por un autómata capaz de identificarlas a partir del contenido de un texto previamente etiquetado. Cualquiera que sea su desarrollo, las marcas de codificación funcionarán como instructor y catalizador para los programas encargados de la generación o interpretación automática.

Como es obvio, la primera opción para la elaboración de patrones no es recomendable porque presenta, ineludiblemente, al menos dos inconvenientes. Por un lado, esta práctica manual es muy onerosa en términos de trabajo humano. Por otro, resulta falible; es más que probable que el error humano se presente en este proceso de reescritura de errores y dé lugar a ausencias o erratas; el lingüista, humano al fin, no contempla fehacientemente todos los componentes que forman parte del elenco de cada paradigma.

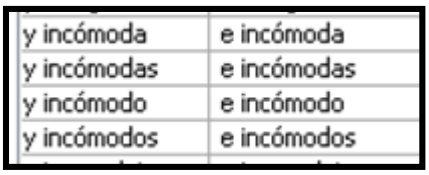
Una primera aproximación a los planteamientos de la segunda opción, —elaboración de patrones codificados— obliga a pensar en la creación de un sistema entendido como un conjunto de herramientas coordinadas que permitan, a partir de unas directrices básicas, tanto generar automáticamente toda la casuística de errores que pueda ofrecer un fenómeno, como interpretar directamente los patrones codificados en el texto etiquetado.

La generación masiva de errores, que deberán alojarse en formas de patrón en algún lugar accesible para el autómata que lleve a cabo la identificación, será

desarrollada por un sistema de recursos que se activarán ante un mismo estímulo; las marcas de codificación que constituyen un lenguaje común entre el lingüista, el programador y la máquina. Estos códigos que se adhieren a algunas formas serán necesarios tanto para el lingüista, que encontrará en ellos un modo expresivo y económico de expresar reglas generales y abstractas, como para el programador, que de manera unívoca llevará a término las ejecuciones que están implícitas en la codificación, en cada marca y etiqueta. Este repertorio de marcas y etiquetas, junto con un formato rígido y una sintaxis precisa conforman el lenguaje común que le permitirá al sistema procesar de forma "inteligente" los patrones y tratar y manipular los datos que estos recogen.

Las acciones que deberá emprender el sistema de generación automática que proponemos se concretan en la flexión, conjugación y listado/desarrollo léxico del paradigma codificado en el patrón que presente marcas.

En el primero de los casos, el FLEXIONADOR deberá replicar el patrón generando errores derivados que surgen de las posibilidades flexivas que presenta una forma canónica. De este modo, un caso como el que se ofreció con relación al corrector de *Word*,



y incómoda	e incómoda
y incómodas	e incómodas
y incómodo	e incómodo
y incómodos	e incómodos

podría sintetizarse del siguiente modo:

Patrón del error	**Corrección**
`y [incómodo/a/os/as]`	`Sustitúyase y por e`

Planteado en términos computacionales, esto es, accesibles para la máquina, el patrón debería utilizar una marca asociada a la forma *incómodo* que actuará como propulsor del flexionador que habrá de llevar a cabo la generación automática de errores agotando todas las formas posibles contenidas en el lema que contiene la marca.

Otro de los recursos en los que se apoya PatErr es un CONJUGADOR, que del mismo modo en que lo hace el flexionador, interpretará las marcas de codificación devolviendo una acción de generación automática de errores. Esta generación, como puede deducirse, solo se aplicará sobre las formas simples y las formas no personales que restan, es decir, gerundio y participio. Generadas estas, quedarán cubiertas todas las formas que pueden formar parte de un patrón de error[74].

En otros casos, el patrón no presentará formas léxicas sino categorías gramaticales necesarias para la expresión genérica y abstracta de estructuras erróneas[75]. Será necesario, pues, crear un subprograma que llamaremos *LEXIFICADOR* cuyo cometido sea interpretar y generar estructuras a partir de etiquetas de categorías gramaticales. El lexificador encarnará con contenido léxico las etiquetas categoriales que aparecen explícitas en el patrón. Retomemos la estructura ya clásica de «*a* + participio»:

a ido	ha ido
a imprimido	ha impreso
a llamado	ha llamado
a pensado	ha pensado
a reído	ha reído
a roto	ha roto

que podría enunciarse de la siguiente manera:

Estructura errónea
a + participio

Una vez más, será necesario asociar una marca a la etiqueta de categoría gramatical que deba ser descodificada para activar el proceso de generación automática o interpretación sobre el texto que habrá de ejecutar el lexificador.

[74] Se generan así, a partir de una forma canónica, un infinitivo 77 formas flexionadas y otras dos para gerundio y participio. Las formas compuestas del verbo serán identificadas en el texto a partir de la forma de participio.

[75] Las categorías gramaticales que se presentan en el corpus responderán al esquema de anotación y especificaciones que se han desarrollado para el Lexicón TIP. Para la correcta explotación del corpus a partir de otros lexicones, será necesario, cuando menos, mantener el planteamiento y repertorio de la categoría funcional de los determinantes.

El último recurso que integra este sistema de representación y generación automática exponencial se basa en un REPERTORIO DE LISTAS de formas que han sido agrupadas por contener alguna característica común. Estas formas generalmente comparten eje paradigmático por lo que son, a nivel estructural, intercambiables.

Las listas actúan del mismo modo en que lo hacen las categorías gramaticales pudiendo incorporarse en el contenido del patrón. El mismo lexificador se encargará de desarrollar la etiqueta que representa a una lista y adscribirá al mismo patrón a todas las palabras contenidas en ella. La naturaleza de las listas es variada y pueden recoger formas correctas o errores creados artificialmente. Para su desarrollo se han aplicado tanto criterios semánticos —expresiones de pasado—, como gramaticales —verbos en 2º persona de singular de pretérito indefinido de indicativo a los que se le añade *-s* final para generar un error—, como morfológicos —sustantivos invariables en plural—. La recopilación de sus elementos se ha llevado a cabo bien manualmente, como es el caso de la lista que contiene sustantivos referidos a las partes del cuerpo, bien de un modo automático, como ocurre con la citada lista de formas verbales. En el Anexo 8 se presentan los títulos de las listas que se han creado. Como cabe esperar, algunos de estos repertorios estarán cerrados mientras que otros, como el destinado a agrupar las siglas, está en constante crecimiento.

Un ejemplo intuitivo que ilustra el alcance del planteamiento que proponemos lo constituye el error provocado por la adición errónea de una *-s* final en las formas del pretérito indefinido de indicativo de segunda persona del singular; **leístes*. Esta epéntesis que quizá surja por analogía con el resto de los tiempos verbales en segunda persona de singular —*lees, leerías, leerás*, etc.—, se encuentra tanto en el español meridional, como en las variedades del español de América; **Tú leístes; *Vos leístes*. Este fenómeno puede suceder con todos los verbos que presenten esta flexión, tiempo y modo verbal, por lo que PatErr deberá contener un patrón de error para cada verbo. Para abordar esta incidencia puede proponerse un patrón como el siguiente:

Patrón codificado	**Corrección**
Lista 35 (todos los verbos en 2ª pers. sg. de pretérito indefinido de indicativo + *s*)	Sustitúyase *-stes* por *-ste*

De este modo, mediante la creación de un solo patrón y de una lista de desarrollada automáticamente —Lista 35: 2ª persona del singular del pretérito indefinido de indicativo de todos los verbos + *s*—, se generarán casi 13 000 errores del patrón codificado. Todas estas derivaciones adoptarán la misma solución que se propone en el patrón codificado.

No obstante, deberá tenerse en cuenta que, formas como la asociada al verbo *ver; *vistes,* no deberán ser corregidas automáticamente debido a que tienen un correlato homógrafo correcto, en este caso una forma de presente del verbo *vestir*. Será necesario, entonces, identificar y aislar este tipo de formas ambiguas, para que el programa no devuelva falsos positivos. Esta restricción y limitación es una consecuencia de los *límites que impone el tratamiento automático*. La máquina es incapaz de hacer una interpretación semántica satisfactoria que resuelva las ambigüedades que cualquier hablante es capaz de discernir sin ningún esfuerzo.

Como puede adivinarse, las cuatro ejecuciones que componen el sistema integrado de generación artificial —flexión, conjugación, lexificación y desarrollo de listas— pueden convivir en un mismo patrón. Así, un patrón que codifique el fenómeno que se esconde tras la estructura «*opción a +* infinitivo» puede, mediante las marcas formales que serán interpretadas por estos subprogramas, por un lado, impeler al flexionador a desarrollar todas las posibilidades flexivas de la forma *opción —opción, opciones, opcioncitas*, etc.— y, por otro, activar el lexificador para que genere patrones sin marcas en las que se sustituya el infinitivo por todas las formas verbales infinitivas que recoja el lexicón del que se nutra. Una vez aplicado el lexificador, el resultado de la generación será un conjunto de patrones que recogerá todas las formas infinitivas del lexicón:

Patrones de error	
`opción a comer`	`opciones a comer`
`opción a hablar`	`opciones a hablar`
`opción a ir`	`opciones a ir`
`opción a tener`	`opciones a tener`
`opción a …`	`opciones a …`

El proceso de descodificación —ejecutado por un algoritmo— derivará, por lo tanto, en la generación automática de toda la casuística posible contenida en una forma o estructura errónea.

El volumen de la generación automática de errores crecerá exponencialmente a medida que se combinan varias etiquetas que contengan un volumen significativo de lemas —por ejemplo en estructuras del tipo «adj + N»— por lo que en ocasiones será conveniente no desarrollar el patrón y tratarlo a tiempo real partiendo de la etiquetación del texto por parte del lematizador.

6.3 Metodología

Para la consecución de los planteamientos que se han expuesto será necesario desarrollar ciertas tareas propias de la ingeniería lingüística en este entorno de lingüística de errores con fines computacionales. Estas tareas, que finalmente se actualizarán en PatErr, se insertan en un proceso que culmina con la elaboración de los patrones formalizados y codificados.

6.3.1 Selección del contenido de los patrones

Para darle contenido al repositorio, es necesario, en primer lugar, llevar a cabo una compilación de la bibliografía y fuentes que tengan por asunto el tratamiento de los ERRORES Y DUDAS MÁS FRECUENTES del español panhispánico.

Se apela a la frecuencia, como ya ha quedado de manifiesto, como criterio para la selección del contenido que se pretende abordar en forma de patrones. De este modo se ofrecerá una cobertura realista que, cuando menos, provea tratamiento para los escollos más comunes y problemáticos que plantea nuestra lengua.

El estudio de frecuencias de los errores del español actual excede las pretensiones de nuestro trabajo, por lo que los resultados de esta investigación previa y necesaria se han confiado a la bibliografía especializada dedicada a estas cuestiones. La selección inicial contó con las siguientes obras de referencia:

- *El libro del español correcto: Claves para hablar y escribir bien en español*. Paredes García, F. (2014).
- *Las 500 dudas más frecuentes del español*; Instituto Cervantes, (2013).
- *Eso no se escribe así: Los 1000 errores más frecuentes en español*. Escarpanter, J. (1993).
- *Dudas y errores de lenguaje*. Martínez de Sousa, J. (1992).
- [Blog de lengua] http://blog.lengua—e.com/. Bustos, A.

Complementariamente, se han extraído fenómenos, excepciones y casos concretos de las obras normativas publicadas por la Academia.

El criterio de extracción se ha basado en la selección de los casos, independientemente de la naturaleza del error y el nivel lingüístico afectado, que tenían presencia en varias de estas obras, esto es, los más frecuentes. Se ha partido del caso particular y a partir del Corpus TIP se ha llevado a cabo un estudio de toda la casuística posible relacionada con el fenómeno en cuestión. Tras acotar todo el espectro del error, en los casos en los que ha sido posible se han planteado extrapolaciones y generalizaciones lingüísticas.

6.3.2 Estudio del fenómeno y acotación de la casuística

Establecida la batería inicial de fuentes, debe procederse a la recopilación y estudio preliminar de los errores que serán objeto de tratamiento. Este estudio, como puede intuirse, va más allá de la asunción de la información gramatical general y, con frecuencia ocasional, que aportan las obras de referencia.

Codificar los errores más frecuentes, implica hacer un análisis minucioso de cada uno de ellos, observando todas las posibilidades, límites y formas que ese error pueda adoptar. Este análisis pormenorizado del fenómeno permitirá, por un lado, elaborar reglas generales que contemplen y anticipen la aparición de los casos, y por otro, cerrar listas, en el sentido de dejar previstos todos los casos en los que se ha comprobado, a partir de las investigaciones apoyadas en el Corpus TIP, que puede surgir el error. Cerrar listas supone que, donde las obras de referencia enuncian los fenómenos, aportan ejemplos y concluyen con un *etcétera*, el lingüista computacional debe identificar todas las posibilidades que se recojan

en ese *etcétera*, para agotar todas las circunstancias en las que puede surgir el error. Pongamos un caso concreto para ilustrar el enfoque que debe adoptarse.

> El dequeísmo aparece a menudo con verbos que se construyen con el sujeto oracional pospuesto, como *preocupar, gustar, apetecer, interesar, fascinar*, etc. (500 dudas: 310).

Para abordar un fenómeno como el dequeísmo se han codificado, en primera instancia, aquellos errores más frecuentes en el uso actual de la lengua que están recogidos en las obras de referencia; **pienso de que, *creo de que, *opino de que,* etc.

Pero esta codificación no extingue ni soluciona el resto de los casos que contienen este error; será necesario, por tanto, un estudio lingüístico a partir del corpus de frases sobre el tipo y naturaleza de los verbos que se prestan a este fenómeno, los contextos comunicativos en los que surgen, las zonas geográficas —si es que existen— en las que se desarrollan con mayor frecuencia o el perfil típico de los usuarios que cometen este error.

Tras este estudio detallado sobre la panorámica impregnada por el dequeísmo, el depósito de errores aumentará su volumen y será capaz de abordar no solo los casos referenciados en las obras consultadas, sino también otros casos residuales, como **reconozco de que, *recuerdo de que, *me desagrada de que*.

Por otro lado, no puede eludirse que, las conclusiones de este estudio deben desarrollarse en forma de patrones computacionalmente tratables y formalizados, que deben observar tanto la teoría y el rigor lingüístico, como las limitaciones que el tratamiento automático impone.

6.3.3 La viabilidad de los patrones: formalización y codificación

Agotada la parte teórica e investigativa consagrada a la recopilación y estudio de lo agramatical e incorrecto se procederá a la formalización y codificación del fenómeno objeto de tratamiento.

Para el enfoque de la lingüística de errores con fines computacionales que debe asumirse, la información bibliográfica de la que se parte estará sesgada y

limitada por la viabilidad en términos computacionales de un sistema como el que proponemos. Un patrón de error debe contener información manejable por un autómata que trabajará con etiquetas gramaticales y rasgos morfológicos y semánticos, por lo que abstracciones como *sintagma nominal*, *oración de relativo*, *complemento directo* o *expresión de la duda* son rasgos/datos que se escapan a sus capacidades interpretativas.

Ante estas coerciones que impone la propia naturaleza del planteamiento y los recursos de los que emerge una herramienta como PatErr, el resultado del estudio previo a la codificación deberá derivar en un patrón o conjunto de patrones viables en términos de formalización y codificación, que den una solución —corrección del error—, respetando no solo los principios de la gramática normativa, sino las limitaciones que impone el entorno del procesamiento automático, en el que además, planean dos fenómenos recurrentemente; la incapacidad interpretativa de la máquina y, en consecuencia, la ambigüedad.

La viabilidad del patrón, por lo tanto, deberá estar contrastada tanto en términos de formalización, como en términos de rigor lingüístico. Para asegurar la validez en términos gramaticales y garantizar que el sistema operará adecuadamente sin generar, por ejemplo, falsos positivos o correcciones indebidas, es necesario contrastar la validez de los patrones diseñados en un corpus, con el fin de verificar que la secuencia que representa el patrón cuando está integrada en un texto es siempre errónea. Por otro lado, el cotejo con el corpus desvelará si hay algún nuevo caso no previsto desde la teoría que obligue a ampliar o matizar el patrón codificado.

Para esta fase de verificación de patrones, los programadores de TIP desarrollaron un subprograma *ad hoc,* BúsquedasCorpus TIP, que permite crear búsquedas selectivas por palabras o categorías gramaticales y aplicarlas al Corpus TIP. Para el ejemplo que estamos tratando, el programa permite rastrear el cotexto previo y contiguo de *de que*, tanto en un universo acotado mediante etiquetas de categorías gramaticales, como en un universo ilimitado sin restringir la búsqueda. El programa devolverá todos los casos, según el formato de búsqueda elegido, compilados de mil en mil, que recogerán una parte significativa de toda la casuística posible que se ha constatado en el español escrito.

Los resultados de BúsquedasCorpus TIP ofrecen un panorama constituido por miles de ejemplos de uso real de la lengua que son necesarios tanto para acotar el alcance de un fenómeno, como para corroborar las hipótesis y generalizaciones de las que se parten antes de formalizar una estructura en forma de patrón codificado. El cotejo de los cotextos servirá, pues, para modelar los patrones y comprobar la viabilidad y certidumbre del contenido que codifican.

Junto con las dificultades previsibles que surgen en estas tareas de diseño, formalización y codificación, deben observarse otros casos de diversa índole que limitan el poder representativo de los patrones; aquellos en los que se aborde una regla lingüística que comporte excepciones, aquellos en los que esté involucrada cualquier forma de ambigüedad, aquellos que no sean errores absolutos y emerjan solo en determinados contextos o cotextos, etc. Todos estos condicionantes deberán ser estimados en los procesos de desarrollo de estas tareas.

6.3.4 Tratamiento del error

Una vez codificado el patrón, que idealmente atrape todas las incidencias, dudas o errores que surjan de un mismo fenómeno, será necesario buscar una corrección, un tratamiento, que de manera automática se ejecute sobre el texto.

La estructuración inicial de toda la casuística alojada en PatErr escinde los *errores* de las *recomendaciones* y de los *avisos lingüísticos* según sea el tipo y la naturaleza de la incidencia que se pretenda tratar. En el Capítulo 8 se abordará esta taxonomía con detalle.

En cualquier caso, el desarrollo de las correcciones tanto para los errores, como para las recomendaciones y los avisos lingüísticos se ha basado, como ya se anunció, en la información y criterios expuestos por las obras académicas. A partir de las orientaciones normativas se procederá a la construcción de soluciones, que, a nivel computacional, se basan en un mecanismo simple en extremo; la extracción del texto original de la secuencia errónea —su extensión puede oscilar desde una letra hasta toda una frase— y su suplantación por una nueva secuencia con la versión corregida o recomendada.

El tratamiento que cada uno de los bloques de incidencia será diferente en términos de automatización. Así, mientras las incidencias registradas con la etiqueta de error podrán automatizar la corrección asociada al patrón, las recomendaciones ofrecerán la libertad de automatizar sus resultados para mejorar el texto. En el caso de los avisos lingüísticos, cuyo contenido e instrucciones están sujetas a una interpretación inteligente del texto, no podrán automatizar este proceso correctivo.

De este modo, ante la identificación en el texto objeto de revisión de las formas derivadas de *barajar*, se activará un aviso lingüístico que informará al usuario sobre la naturaleza semántica de este verbo, que no debe emplearse con el sentido de *considerar* cuando se hace referencia a una sola cosa u opción; **Estoy barajando la opción de irme a las misiones*. Para este caso, que requiere una interpretación inteligente sobre la naturaleza —plural o singular— del complemento directo que acompaña a *barajar*, el programa solo podrá advertir al escritor sobre esta restricción semántica y ofrecer alternativas a esta forma; *estudiar, analizar, considerar, proponer,* etc. Será el escritor quien, finalmente, adopte la opción correctiva que considere más adecuada a partir de la información proporcionada por el programa.

La tabla que acompaña a estas líneas presenta sintéticamente las posibilidades de automatización del tratamiento que se propone en PatErr para cada patrón de error.

ERROR	Corrección; automática
RECOMENDACIÓN	Recomendación; corrección automatizable o interactiva
AVISO	Sugerencia, advertencia; corrección interactiva

Las medidas correctivas que se han desarrollado para el tratamiento de las incidencias han supuesto para la tarea lingüística retos de diversa magnitud; mientras algunos errores relacionados con los niveles ortográfico o morfológico permiten establecer generalizaciones —macrorreglas—, que solventan de modo más o menos sencillo bloques enteros de fenómenos, habrá otros de carácter más abstracto que requieran un estudio pormenorizado del fenómeno, de su contexto y cotexto y de su posible ambigüedad.

6.3.5 Anotación de otros rasgos

Como ya se ha expuesto, PatErr está ideado y desarrollado para ofrecer cobertura e información sobre el error trascendiendo a la clásica corrección automática; junto a las etiquetas relacionadas con las incidencias que se detectan en el texto —error, recomendación y aviso lingüístico—, el recurso que se presenta ofrece otras informaciones relativas al nivel lingüístico o área afectada, a la variedad lingüística en términos de variedad diatópica, diafásica y diastrática, el registro y la referencia normativa que avala el tratamiento llevado a cabo por el programa.

Por otro lado, como se ha señalado en páginas precedentes, un complemento importante que aporta este trabajo a otros revisores textuales es la información, en forma de glosa, que se mostrará al usuario asociada a cada situación lingüística que se haya tratado. Para ello se han reescrito las descripciones gramaticales de las obras de referencia adaptándolas a un usuario hipotético representativo. Las glosas, junto con los ejemplos de uso y las referencias bibliográficas, conforman el servicio básico de asesoría lingüística que ofrece el recurso que aquí presentamos.

6.4 Formalización del error: los patrones

Conviene aclarar, en este punto avanzado de la exposición[76], que cuando se hace referencia a un patrón de error, se alude estrictamente al registro formalizado y generalmente codificado —con marcas— de una secuencia o estructura que contiene una incidencia, ya se desarrolle como un error propiamente dicho, una recomendación o un aviso lingüístico.

Como se ha expuesto, cada patrón debería encapsular todas las posibilidades de error que se pudieran generar asociadas al fenómeno lingüístico que intenta subsanar pero, como puede aventurarse, la realidad y los azares lingüísticos impiden pensar en un sistema con ese grado de eficacia.

[76] A partir de los siguientes capítulos, la base de datos donde se asienta el corpus cobrará importancia. Este alojamiento contiene, entre otros campos, uno —de 18— cuya etiqueta es patrón de error. Será necesario, por lo tanto, acotar y precisar esta expresión y ubicar su posición en el sistema al que pertenece.

Hay casos, los menos, y generalmente relacionados con temas ortográficos[77], en los que sí es posible establecer una correspondencia unívoca entre una incidencia y su corrección, a partir de un solo patrón de error codificado, pero estos son casos aislados. Un ejemplo simplificado de estos patrones de aplicación generalizada puede ser el ya propuesto para subsanar formas como **escuchastes.*

Patrón codificado	**Corrección**
Lista 35 (todos los verbos en 2ª pers. sg. de pretérito indefinido de indicativo + *s*)	Sustitúyase *-stes* por *-ste*

En este caso un mismo patrón puede, por sí solo, cubrir todos los contextos que pueda conquistar un fenómeno, la epéntesis verbal de *-s*, en este caso.

Pero hay otra gran parte de fenómenos o errores, que no se pueden tratar absolutamente a partir de un solo patrón y será necesaria la concurrencia de varios patrones, una batería, que parcialmente vayan solucionando partes de todo el espectro de errores que el fenómeno genera. Intuitivamente, no es difícil adivinar que no existe la posibilidad, bajo la propuesta que aquí se desarrolla, de elaborar un patrón único capaz de solucionar todos los errores relacionados con la dupla *sino/ si no,* el uso del gerundio o el tratamiento de ciertas concordancias que entrañan dificultad para el usuario de español.

Por otro lado, habrá casos en los que el patrón no exija marcas de codificación y podrá ofrecer el tratamiento previsto directamente, sin la intervención de las herramientas auxiliares;

Patrón de error	**Corrección**
a groso modo	Sustitúyase *a groso modo* por *grosso modo*
a grosso modo	Sustitúyase *a grosso modo* por *grosso modo*

[77] Nos referimos aquí a errores de ortografía que se adscriben a una palabra, a diferencia de errores provocados por temas de homofonía que afectan a más términos; *a ver/ haber*.

De esta breve panorámica se desprende que, en PatErr están acogidos manteniendo el mismo formato y tratamiento, tanto los patrones codificados, que representan un elenco de errores en potencia a partir de una misma configuración, como patrones de error exentos de marcas, que sí se representan a sí mismos como errores.

6.4.1 Patrones codificados

Estos patrones son introducidos manualmente por el lingüista en la base de datos y son producto de la formalización de un error o una estructura errónea.

Los patrones codificados deben ser lo más genéricos posible, es decir, abarcar el máximo espectro del fenómeno que están tratando. Para cumplir esta premisa, será necesario que cuando el contenido de un error requiera la formalización de configuraciones sintácticas, estas sean abarcadoras y presenten su proyección expandida. Para la codificación de un patrón que intente ofrecer solución al error de segmentación que se observa en oraciones como **Estuvimos en un en torno maravilloso,* será necesario partir de una estructura del tipo «det + *en torno/os».* Con este patrón quedarían excluidas oraciones como **Volvería a aquel maravilloso en torno*[78]*.* Será necesario, por lo tanto, habilitar un lugar en la estructura sintáctica del sintagma nominal para dar cabida a la posibilidad de que haya un adjetivo que complemente al sustantivo **en torno*[79].

La combinatoria de las posibilidades configuracionales que presenta un sintagma nominal es, como se sabe, inagotable en cuanto a la variedad de unidades anidadas que puede acoger, pero partiendo de la asunción de que lo que se

[78] Para ofrecer un dato sobre el potencial en términos de cobertura del corpus que presentamos puede observarse que el corrector de *Microsoft Word* 2016 solo es capaz de señalar como posible error el presente en la primera oración.
Otros correctores disponibles en la red como *Stilus*, *SpanishChecker* y *LanguageTool*, no reconocen error en ninguna de las oraciones.
http://spanishchecker.com/
http://www.mystilus.com/Correccion_interactiva
https://languagetool.org/es/

[79] Como se observará, este criterio de ampliar las proyecciones se aplicará, en numerosas ocasiones en el entorno verbal, posibilitando que un adverbio que modifique al verbo entre a formar parte de las configuraciones verbales. Un caso como **Entiendo absolutamente de que quieras acompañarle,* en el que se intercala un adverbio entre la forma verbal y *de que*, necesitará, para ser tratado, que la estructura que se presente en el patrón sea capaz de acoger a este complemento.

pretende solucionar mediante este sistema son errores en contextos contiguos será necesario llegar a una solución de compromiso en el diseño de cada patrón. Ampliar excesivamente las posibilidades de proyección del sintagma, en el caso ilustrado formalizando para la máquina, por ejemplo, una oración de relativo, puede llegar a sobredimensionar un error que, en términos generales, se ofrece en un contexto más escueto.

Por otro lado, la cantidad de errores que puede derivar de un mismo patrón codificado como resultado de la combinación de todas las formas posibles de cada paradigma que intervenga en el patrón puede alcanzar cotas abrumadoras llegando a saturar el alojamiento de los patrones con errores improbables o imposibles o pudiendo comprometer la velocidad de ejecución del programa si se plantea el proceso de revisión textual a tiempo real.

La solución que se ha adoptado para este caso pretende encontrar el equilibrio entre la legibilidad y el realismo que debe representar todo patrón y la cobertura que este pueda ofrecer.

Patrón de error	**Corrección**
`det + posibilidad de adj + *en torno*`	`Sustitúyase *en torno* por *entorno*`

Como se ha observado, en el interior de un patrón codificado se pueden combinar tanto palabras como marcas de codificación o categorías gramaticales, cuyos límites y componentes serán los expuestos y derivados del etiquetario desarrollado para el Lexicón TIP. Estas marcas deben ser computacionables, por un lado, para la explotación y maximización de la casuística derivada de la configuración expresada en el patrón, y por otro, para la interpretación de los códigos y su identificación sobre el texto etiquetado en una ejecución de identificación directa entre patrón de error y el contenido del texto.

Estas dos opciones, generación automática del patrón e interpretación del contenido abstracto sobre el texto, podrán ejecutarse sobre un mismo patrón, esto es, pueden generarse patrones que deriven de uno codificado que no desarrollen todas las etiquetas que contiene. Para el caso que presentamos se generarían

patrones que tuvieran actualizado el contenido de la etiqueta «det» pero dejaran sin lexificar «adj»]. De este modo quedarían patrones del tipo:

el «adj» en torno
nuestro «adj» en torno
aquel «adj» en torno
tal «adj» en torno

cuyo destino sería el cotejo directo sobre el texto.

En cualquier caso, el planteamiento y desarrollo de PatErr permite la aplicación de los patrones con un amplio grado de flexibilidad toda vez que permite explotar el sistema con las mismas herramientas y marcas de codificación. Será el desarrollador del autómata, condicionado por el modo en que quiera llevar a cabo la revisión —sobre datos prefabricados o sobre estructuras que se identifican y desarrollan a tiempo real— el que opte por una u otra vía de planteamiento, bien la generación masiva de todo el patrón, bien por la generación controlada de alguna de sus partes.

6.4.1.1 *Orden de prelación de los patrones*

El orden de aplicación de los patrones se organizará según los niveles lingüísticos objeto de intervención.

La ejecución del programa que se nutra de PatErr debe seguir, en síntesis, un proceso de rastreo sobre el texto que desee revisarse en busca de fragmentos de texto que repliquen el contenido real o figurado que contienen los patrones.

Una vez sea detectada e identificada una secuencia tratable, deberá observarse si contiene un error de naturaleza ortográfica. Si es así, se tratará esta incidencia y se rastreará el texto de nuevo. Así una secuencia errónea como **Intullo de que vendrá*, no estará recogida, en principio, como un patrón de error. Será necesario corregir **intullo* por *intuyo* para que el sistema encuentre correlato en el repositorio de errores y proceda a su tratamiento: **Intuyo de que - Intuyo que.*

Tras el reanálisis posterior al tratamiento ortográfico, se procederá a la intervención de los otros niveles, recuperando de la base de datos de error toda la información asociada para su tratamiento.

Por otro lado, en cuanto a la escisión que se propone entre errores, recomendaciones y avisos, cabe establecer un orden subsidiario para evitar interferencias entre patrones de distinta naturaleza —error o aviso— pero mismo contenido —tratamiento de secuencias erróneas en las que participe *por qué* o *porque*—. Los primeros que deberán ejecutarse, una vez identificada una forma en el texto, serán los errores y las recomendaciones. El contenido de unos y otros no se solapa, por lo que es posible hacer una ejecución unificada de estos patrones. Los avisos, que en muchas ocasiones contienen formas a las que no puede ofrecérsele un tratamiento que subsane el posible error, solo se activarán como último recurso en un análisis final con el objetivo de ofrecer información relevante al escritor para aquellos casos susceptibles de contener un error que no han recibido un tratamiento previo.

Capítulo 7

Formalización de los patrones de error. Lenguaje de codificación

Uno de los cambios de perspectiva más notables que tiene que afrontar el lingüista teórico tradicional en este nuevo marco de la ingeniería lingüística se materializa en el lenguaje que debe adoptar para comunicarse con éxito con el informático y, en última instancia, con la máquina. Es, pues, necesario fijar el modo de codificación y de formalización de la teorización lingüística, de modo que los datos y reglas codificadas sean accesibles, manipulables y tratables por las aplicaciones informáticas.

Como se dijo al comienzo de este trabajo, *solo lo formalizable es programable.* Ante este *a priori*, una de las primeras tareas es consensuar un lenguaje que satisfaga a las dos partes —lingüística y computación— y observe tanto las limitaciones como las capacidades que ambos ámbitos presentan.

Será necesario que el lingüista emprenda una adaptación y transferencia de sus investigaciones y propuestas elaboradas para dar solución —corrección y asesoramiento— a los problemas —errores— dentro del nuevo escenario computacional de la ingeniería lingüística. Del mismo modo, es cada vez más productivo que el ingeniero informático conozca las unidades de análisis gramatical, así como las relaciones que se establecen entre ellas dentro del sistema de la lengua.

Entre las capacidades más productivas que presenta la máquina destacan el cómputo de datos y la ejecución automática masiva de órdenes precisas y procesables en su entorno. Estas capacidades extraordinarias de generación y cálculo de máxima precisión en tiempos exiguos —en comparación con el tiempo y coste que le suponen este tipo de tareas al humano— son algunas de las potencias

que permiten materializar los planteamientos y programas diseñados en el entorno del PLN.

Dentro de las limitaciones que presentamos los seres humanos se encuentra la incapacidad de acceder introspectivamente a todos los datos que disponemos. Un hablante competente es capaz de advertir una secuencia errónea como **No habemos comido todavía*, pero si se le pregunta por el número de posibilidades erróneas que puede presentar esta estructura, posiblemente dude, yerre, y se muestre incapaz de ofrecer un listado lo suficientemente representativo con respecto al número exacto de casos que puedan derivar de este fenómeno. A pesar de esta insuficiencia, sería capaz de identificar y catalogar como erróneas todas esas secuencias derivadas.

Uno de los posibles puentes que se pueden crear para aunar ambas capacidades es, en el entorno en el que trabajamos, el establecimiento y uso de un lenguaje de codificación común que permita al lingüista crear generalizaciones, *condiciones de error,* a partir de un lenguaje expresivo que, interpretado por la máquina, provocará la generación automática o identificación en el texto de todos los casos posibles que se incluyan en esa generalización.

La incorporación de estas marcas supone una novedad con respecto a otros patrones de error desarrollados de los que se tiene constancia[80] y permite crear generalizaciones codificadas que, interpretadas por la pléyade de subprogramas que hemos presentado —flexionador, lexificador, etc.— podrán emprender un proceso expansivo de generación automática de errores.

7.1 Las marcas de codificación

Las convenciones y especificaciones relativas a la codificación de los patrones de error que se han adoptado responden a la naturaleza del planteamiento y recursos que lo constituyen. Por este motivo se adoptará, en lo que respecta a las categorías gramaticales, la fundamentación lingüística que se expuso para el Lexicón TIP junto con las abreviaturas que sirvieron para su formalización y codificación. En el

[80] Nos referimos, una vez más, a parte del corrector instalado en el procesador de *Microsoft Word*.

Anexo 9 puede consultarse la lista de las abreviaturas contenidas en los patrones de PatErr.

La metodología planteada para la codificación de los patrones está inspirada en las EXPRESIONES REGULARES de la ciencia computacional teórica y de la teoría del lenguaje formal. Se considera que una expresión regular es una *secuencia de caracteres y metacaracteres*[81] *que forma un patrón de búsqueda*, que normalmente se aplicará sobre cadenas de caracteres. Las expresiones regulares ofrecen una gran flexibilidad para llevar a cabo la tarea de codificación y modelización del conocimiento lingüístico a partir de una sintaxis simple y lineal que permite crear estructuras complejas donde pueden combinarse tanto contenido literal, como marcas figurativas, —metacaracteres—. En palabras de Friedl:

> It might help to consider regular expressions as their own language, with literal text actig as the words and metacharacter as the grammar. The words are combined with grammar according to a set of rules to create an expression that comunicates an idea (2005:5).

Aunque el planteamiento y enfoque adoptado para este trabajo es el mismo, los metacaracteres utilizados no coinciden con los estandarizados debido a que las necesidades lingüísticas de esta propuesta no emergen en el ámbito de la ciencia computacional teórica.

A continuación, se presenta la lista de marcas empleadas para la codificación tanto de los patrones de error, como de su tratamiento:

PATRONES	
Marca	**Descripción**
{}	Listar
#	Conjugar
##	Restringir la conjugación
@	Flexionar
@@	Restringir la flexión
^	Exceptuar
?	Ninguna o una vez

[81] Entendemos por *metacarácter* un carácter del propio lenguaje que incorpora una funcionalidad adicional en la búsqueda del patrón más allá de su propia representación, esto es, su significado no es literal.

+	Una o más veces
\w	Cualquier palabra

TRATAMIENTO	
Marca	**Descripción**
[ ]	Contiene caracteres
+	Añadir
-	Eliminar
|	Operador OR, uno u otro

7.1.1 Las llaves «{ }»

Los metacaracteres «{ }» indican que el contenido que acogen está codificado, por lo tanto, no se debe buscar su literalidad en el texto objeto de escrutinio. Este contenido puede concretarse en categorías gramaticales, rasgos morfosintácticos, rasgos semánticos o una lista. Algunos ejemplos de uso en este trabajo:

```
{inf}; {N}; {subj}; {pl}; {reflex}; {locuciones latinas};
                    {cuantificador}
```

Como puede deducirse, el contenido ceñido por estas marcas debe estar etiquetado y compilado en algún lugar —en este caso en el lexicón y en listas prefabricadas—, al que tenga acceso el programa para poder rescatar los elementos que se pretenden incluir en «{ }» . De esta suerte, y ajustándonos a las limitaciones de las herramientas de análisis lingüístico que se utilizan para esta propuesta, no será posible incluir entre estos códigos elementos o configuraciones tales como SN, SP, O. Relativo, etc.

7.1.2 La almohadilla «#»

La marca «#» representa todas las formas conjugadas simples del infinitivo que le sigue. Este código en un patrón permite representar cualquier forma conjugada del verbo al que se asocia [82].

```
#tensar  #comer  #tener  #aprobar
```

Habrá ocasiones, en las que el contenido codificado no requiera una generación o interpretación irrestricta sino precisada y acotada. Para estos casos se ha previsto la doble almohadilla «##» que permita al autómata identificar el tipo de coerciones que debe aplicar. De esta forma, el patrón irrestricto:

```
#soñar
```

permitirá generar cualquier forma conjugada del verbo *soñar*. Sin embargo, el patrón selectivo:

```
#soñar##{subj}##{pl}
```

generará solo las formas plurales del subjuntivo del verbo *soñar*.

7.1.3 La arroba «@»

El metacarácter «@» representa todas las formas flexionadas de la palabra que le sigue. En el caso de los sustantivos, las flexiones previstas serán las propias de género, de número y las derivadas de la flexión a partir de sufijos apreciativos. Las mismas flexiones se prevén para los adjetivos que además presentarán la flexión propia del grado superlativo. Se contempla, además, la flexión de unos pocos adverbios que aceptan marcas flexivas como en *cerquita, prontísimo* o *despacito*. Esta marca, por lo tanto, no se aplicará ni a las preposiciones, ni a las conjunciones.

```
@casa  @amanecer  @blanco  @mejicano @cerca
```

Se han utilizado dos códigos diferentes para representar las formas conjugadas de las flexionadas por dos motivos: por un lado, los recursos encargados de interpretar estas marcas son diferentes —conjugador y

[82] En el elenco de esta representación se incluye, además, el diminutivo del gerundio para aquellos pocos casos que están aceptados por la Academia y son de uso frecuente en algunas regiones latinoamericanas; *corriendito, andandito,* etc.

flexionador— y, por otro, los códigos servirán como mecanismo desambiguador en casos en los que como `amanecer` sean posibles los dos tipos de flexiones, tanto la nominal, como la verbal.

Como sucede con el tratamiento verbal, habrá ocasiones en las que el contenido codificado no requiera una generación irrestricta sino ceñida. Para estos casos se ha previsto la doble arroba «@@» que permita al autómata llevar a cabo estas restricciones. De esta forma, el patrón general

```
@niño
```

permite generar cualquier forma flexionada del sustantivo *niño*. Sin embargo, el patrón selectivo

```
@niño@{fem}@@{pl}
```

generará solo las formas femeninas plurales: *niñas*, *niñitas*, *niñicas*, *niñazas*, etc. Como puede deducirse, la cobertura del flexionador que interprete el metacarácter incidirá directamente sobre las producciones o interpretaciones que pueda llevar a cabo en el proceso de revisión textual.

7.1.4 El acento circunflejo «^»

Otro modo de precisar el contenido de un patrón es excluir de su ámbito de acción algunas formas, rasgos flexivos o etiquetas semánticas. Para ello, se utilizará la marca «^» que indica un *excepto* del contenido incluido entre los paréntesis que le preceden.

```
#cuidar^({part})
```

Este segmento impelerá al conjugador a la generación o representación de todas las formas conjugadas del verbo *cuidar* excepto las participiales.

7.1.5 El interrogante de cierre «?»

El metacarácter «?» indica que la secuencia o etiqueta que le precede es opcional. Su utilización flexibilizará el patrón de error dando cabida a elementos opcionales que suelen concurrir en ciertas estructuras.

```
#estar @{adv}? desecho
```

En este ejemplo de error, el interrogante de cierre indica al autómata que la estructura que genere o identifique puede contener intercalado un adverbio. Así, la interpretación completa de este patrón sería "cualquier forma conjugada del verbo *estar* seguida opcionalmente de un adverbio y la palabra literal *desecho*"

7.1.6 El signo de adición «+»

La marca «+» indica la posibilidad de que haya más de un elemento del mismo tipo del que acompaña en esa posición. En este ejemplo,

```
bajo @{det}+ @planteamiento
```

la interpretación que debe hacerse del signo «+» es "hay al menos un determinante pero puede haber más". En el caso de estas configuraciones, con determinantes las múltiples formas que pudieran presentarse están separadas por espacios en blanco;

**bajo el otro planteamiento*

**bajo todos aquellos otros planteamientos*

Sucede con otras configuraciones con adverbios del tipo **Hecha casi siempre en falta.*

Sin embargo, en otras ocasiones las etiquetas de categorías deberán estar separadas por la coma y/o por conjunciones. En el siguiente ejemplo se observa cómo para el desarrollo de estas estructuras complejas es necesaria la presencia de algunos de estos elementos de coordinación; **Da la absoluta, fatídica y nefasta casualidad de que lo vi demasiado tarde.*

Para que el patrón represente estas estructuras será necesario, por lo tanto, que el signo «+» incluya a estos elementos en su ámbito de aplicación.

```
#hechar @{adv}+ en falta
#dar la @{adj}+ casualidad que
```

Este metacarácter combinado con la marca «?» deberá interpretarse como ausencia o presencia de uno o varios elementos. De esta forma podemos completar la representación de los dos ejemplos anteriores como sigue:

```
#hechar @{adv}+? en falta
#dar la @{adj}+? casualidad que
```

De este modo, el nuevo patrón asumirá tanto la versión simplificada de los ejemplos anteriores **hecha en falta* y **da la casualidad de que* como los que presenten estructuras más complejas; **hecho casi deseperadamente en falta un poco de comprensión*.

Como la marca del apartado anterior, el signo «+» permite atenuar la rigidez de las configuraciones de los patrones codificados, es decir, los flexibiliza y optimiza su capacidad figurativa.

7.1.7 La barra invertida con una letra «\w»

Por último, la marca «\w» informa sobre la posibilidad de que en esa posición el patrón presente cualquier elemento de la lengua, excepto los signos de puntuación que delimitan una oración —[.][?][!]— y el punto y coma.

```
no solo \w si no
```

Es evidente que para ofrecer tratamiento a esta estructura es necesario ampliar la cobertura para que admita varios elementos entre los literales invariables, que no llevan marcas de codificación, con el fin de facilitar la identificación de errores como:

**no solo me miras si no…*
**no solo bebes y fumas si no…*
**no solo me besaste y me acariciaste, eso sí, siempre con respeto, si no…*

De no expandir el ámbito de «\w», el patrón quedaría muy limitado y sin poder de representación. Para alcanzar este efecto y generar patrones realistas, le incorporamos la marca «+» al metacarácter «\w»

```
no solo \w+ si no
```

La delimitación impuesta a la cobertura del código «\w», en cuanto a los signos de puntuación que no contempla, evitará detectar como erróneas estructuras correctas como *No solo quiero que me hagas el desayuno todos los días; si no me invitas a cenar los lunes, no haremos ningún trato.*

7.1.8 Los corchetes «[ ]», el signo menos «-» y el signo más «+»

Los corchetes se emplearán en las soluciones propuestas para cada fenómeno con el fin de indicar los caracteres que deben ser tratados en la estructura detectada en el texto que replica el patrón. Si el contenido del corchete se inicia con el signo menos «-» indicará que se debe eliminar de la estructura del error en el texto todas las secuencias que coincidan con el contenido del corchete. Cuando el contenido del corchete comience con el signo más «+» indicará que se le debe añadir a la estructura detectada los caracteres contenidos en el corchete. De este modo, se llevará a cabo la sustitución del contenido erróneo por otro contenido que, inserto en la cadena tratada, dará lugar a una secuencia correcta.

Patrón	Tratamiento	Texto corregido
`estadunidense`	`[-u][+ou]`	estad**ou**nidense
`el Cairo`	`[-el][+El]`	**El** Cairo

Como se observa en los ejemplos, la solución propuesta provocará que el autómata extraiga el primer carácter «*u*» de la estructura y lo sustituya por la secuencia «*ou*». En el segundo caso buscará la secuencia «*el*» y la sustituirá por «*El*».

La definición de lo que contienen los corchetes está planteada en caracteres y no en palabras porque no siempre será necesario sustituir una palabra completa y, por otro lado, en ciertas ocasiones será necesario contemplar una secuencia de corrección que se extiende más allá de la palabra;

Patrón	Tratamiento	Texto corregido
`de seguido`	`[-de seguido]` `[+ininterrumpidamente]`	ininterrumpidamente
`los unos de las otras`	`[-las otras]` `[+los otros]`	los unos de los otros

Este planteamiento que se adopta para el tratamiento del error evita que el autómata tenga que llevar a cabo análisis de alto nivel de abstracción para los que se requiere algún tipo de analizador sintáctico automático.

7.1.9 La barra vertical «|»

Por último, la marca «|» indica optatividad entre los elementos situados a ambos lados de su posición. Dicho de otro modo, indica una disyuntiva entre varias opciones.

Patrón	Tratamiento
`software`	`[-software]` `[+programas]|[+aplicaciones]`

`a día de hoy`	`[-a día de hoy]`
	`[+hoy en día]|`
	`[+hoy por hoy]|`
	`[+hoy día]|`
	`[+en la actualidad]`

Según estos ejemplos, el tratamiento previsto para la primera recomendación será la sustitución de la forma *software* por una de las opciones sugeridas; *programas* o *aplicaciones.* En la segunda recomendación se ofrecen varias opciones preferidas para expresar la locución *a día de hoy*.

La primera opción que aparezca codificada ocupará esta posición por ser la solución más frecuente o recomendada según la norma. De este modo, para la automatización de estos resultados, será la opción que se presente en primera posición la que sustituya automáticamente a la palabra afectada por el error.

7.1.10 Consideraciones

7.1.10.1 *Múltiples opciones para codificar un mismo error*

Los metacaracteres de las expresiones regulares nos permiten optar, en numerosas ocasiones, por distintas codificaciones y soluciones para abordar un

mismo fenómeno o estructura errónea. Esta pluralidad de opciones surge, por un lado, como consecuencia de la diversidad de enfoques que puede adoptar un lingüista para representar un fenómeno, y por otro, por la citada flexibilidad de las marcas y de la lógica de la codificación que planteamos. En cualquier caso, el criterio que se ha adoptado en el desarrollo de PatErr busca el equilibrio entre legibilidad, realismo y cobertura.

A continuación, se presentan las distintas opciones para la codificación de la estructura *detrás mío/a*. Siguiendo un criterio generalizador se puede codificar en un único patrón como:

```
detrás @mío
```

A pesar de que la marca «@» representa a las cuatro formas del pronombre *mío —mío*, *mía*, *míos* y *mías*— aunque este patrón representa más errores de los reales derivados de la expresión incorrecta **detrás mío/a*, no afectaría ni a la identificación ni a la solución —sustituir *mío/a/os/as* por *de mí*—.

Si optamos por desglosar la estructura errónea en todas sus opciones *detrás mío* y *detrás mía* y le aplicamos un patrón a cada una, tendríamos una batería de patrones para cubrir este fenómeno:

```
detrás mío
detrás mía
```

Y si ajustamos el patrón exactamente para cubrir el espectro real del error tendríamos que codificarlo como:

```
detrás @mío@@{sg}
```

Consideramos que, para un ejemplo como este, se puede aplicar el criterio de máxima generalización sin perder legibilidad ni mermar la eficiencia computacional. Por este motivo, la opción que generalmente se ha elegido para la codificación del contenido de PatErr ha sido esta opción simplificada. No obstante, como ha quedado demostrado, los tres planteamientos descritos son viables.

Sin embargo, no siempre todas las posibles opciones de codificación de una estructura errónea cumplen objetivamente los criterios de legibilidad y comprensión como se demuestra en el siguiente ejemplo. La estructura sintáctica «*al contrario de* + SN» podemos codificarla en un único patrón aglutinando todas

las posibles proyecciones que puede ofrecer un sintagma nominal[83] (se utiliza el paréntesis para agrupar las opciones posibles separadas mediante el símbolo «|»):

```
al contrario de (@{pn}|@{det}+ @{adj}+? @{N}|{prep}|@{N}@@{propio})
```

Esta opción parece poco recomendable porque compromete la legibilidad del patrón innecesariamente. En ocasiones, será necesario intervenir sobre un patrón codificado para corregir, matizar o restringir su alcance y ante patrones enmarañados como estos se dificulta la manipulación de las estructuras.

Otra opción posible para la codificación de este patrón sería desglosar al máximo las posibilidades configuracionales de esta estructura sintáctica y generar un patrón para cada una, es decir, una batería de patrones para un mismo fenómeno;

```
al contrario de @{pn}
al contrario de @{prep}
al contrario de @{N}@@{propio}
al contrario de @{det} @{N}
al contrario de @{det} @{det} @{N}
al contrario de @{det} @{det} @{det} @{N}
al contrario de @{det} @{adj} @{N}
al contrario de @{det} @{det} @{adj} @{N}
al contrario de @{det} @{det} @{det} @{adj} @{N}
al contrario de @{det} @{adj} @{adj} @{N}
al contrario de @{det} @{det} @{adj} @{adj} @{N}
al contrario de @{det} @{det} @{det} @{adj} @{adj} @{N}
```

Como se sabe, esta batería de patrones desglosada tampoco recogería todas las combinaciones posibles que pueden registrarse en un sintagma nominal, pero sirve para ilustrar la complejidad que genera la elección de esta opción para llevar a cabo la codificación.

Tal y como se señaló en el capítulo anterior, se recomienda codificar con un criterio que busque el equilibrio entre la legibilidad y la comprensión del patrón con la generalización de la casuística que se persigue, esto es, la cobertura. Por lo

[83] Se excluyen de las proyecciones configurables para un sintagma nominal opciones desarrolladas por componentes como las oraciones de relativo, que impiden la localización del error en un contexto de contigüidad.

tanto, la opción recomendada y elegida para la codificación de este tipo de estructuras en PatErr será la siguiente:

```
al contrario de @{pn}
al contrario de @{det}? @{adj}+? @{N}
al contrario de {prep}
al contrario de @{N}
```

Por otro lado, existen casos en los que la cantidad de patrones para representar una estructura viene obligada por el tipo de correcciones que deban aplicarse.

En la estructura «*el uno* + prep + *la otra*» que representa una recomendación, la corrección dependerá de la preposición utilizada. Así la expresión **los unos hacia la otra* se debe corregir sustituyendo *la otra* por *el otro,* pero cuando la preposición involucrada sea *d*e o *a* la corrección deberá aplicarse con las partículas contractas: *del otro* o *al otro*. Serán necesarios, por lo tanto, tres patrones para codificar la misma estructura:

Patrón	Tratamiento
`el uno {prep}^de^a la otra`	`[-la otra][+el otro]`
`el uno a la otra`	`[-a la otra][+al otro]`
`el uno de la otra`	`[-de la otra][+del otro]`

7.1.10.2 *Las mayúsculas*

En ciencias de la computación el carácter en minúscula es diferente de su correlato en mayúscula y cualquier tratamiento que requiera ser indiferente a este hecho debe solucionarse mediante programación. Todos los buscadores textuales realizan su tarea con independencia del estado ortotipográfico —mayúscula o minúscula— en el que se encuentre la cadena de búsqueda.

Aplicando un criterio similar, las minúsculas en el patrón permitirán buscar e identificar errores en los dos estados de las letras, sin embargo, la mayúscula buscará su correspondiente en el texto únicamente en mayúscula, siendo, por lo tanto, privativa.

Así, las estructuras de error localizadas al principio de la frase —donde la primera letra se presentará en mayúscula— quedarán cubiertas. El siguiente patrón,

```
la minoría de @{N}@@{fem pl}
```

cubre tanto al error inserto en la frase como al ubicado al principio, a pesar de que en el segundo caso la cadena de error del texto comenzará con mayúscula.

**Le gusta a la minoría de niñas*

**La minoría de gatas son mansos.*

Como se observa en el siguiente ejemplo, el uso de la mayúscula sí es excluyente, por lo que su uso indicará indirectamente que el patrón debe encabezar la frase.

```
En primer lugar, decir que
```

Del mismo modo, la secuencia *Toca las palmas muy fuerte* no se verá afectada por el patrón ortotipográfico que corrige *las Palmas* por *Las Palmas* ya que para llevar a cabo esta corrección el texto debe presentar la letra *P* mayúscula:

```
las Palmas
```

7.1.10.3 *La cadena de búsqueda en la secuencia de tratamiento*

Existen varias opciones de formato para codificar las soluciones siguiendo la metodología planteada. En el siguiente ejemplo de aviso lingüístico se ilustran algunas soluciones válidas para su tratamiento:

Patrón	Tratamiento
aún así	[-ú][+u]
	[-aún][+aun]
	aun así

La primera opción plantea una sustitución simplificada tratando únicamente los caracteres erróneos. La segunda solución tiene un enfoque más conceptual al proponer la sustitución de las palabras afectadas. Por último, la tercera solución propone la sustitución completa de la estructura errónea por la correcta. Las dos

primeras son equivalentes en tiempos de computación, puesto que en las dos es necesario realizar operaciones con las cadenas de carácter en la solución y búsquedas en la estructura errónea.

Pero desde el punto de vista computacional, la más eficiente en tiempo de respuesta es la tercera, dado que no es necesario realizar operaciones de análisis de cadenas de caracteres para interpretar y descodificar la solución propuesta y tampoco es necesaria una búsqueda en la estructura errónea localizada en el texto, —se procede a una sustitución completa de una estructura por la otra—.

El criterio aplicado para codificar las soluciones dependerá del peso que el lingüista considere más adecuado de los siguientes factores que deben tenerse en cuenta; *comprensión* y *rendimiento computacional.* Minimizar los errores humanos mientras se manipula PatErr es, finalmente, más importante que el rendimiento computacional que puede suponer la elección de una u otra opción, por ello, lo recomendable es facilitar la comprensión del tratamiento propuesto para el investigador que manipula los datos.

Debe advertirse, con relación a la eficiencia y coste computacionales, que los patrones que integran metacaracteres en su tratamiento son más onerosos en términos de computación que aquellos que se procesan y ejecutan sin marcas.

En algunos casos de codificación se desconoce *a priori* cuál de las estructuras erróneas que cubre será la que se encontrará en el texto. Este hecho, el desconocimiento de la secuencia que el autómata se puede encontrar en el flujo de texto, por ejemplo, bajo la etiqueta {adv}, debe tenerse en consideración para el planteamiento y desarrollo de la corrección. Así un patrón como el siguiente,

```
hecha a {inf}
```

no puede ser sustituido por una estructura completa decidida de antemano, debido a que no se conoce el verbo que formará parte de la estructura errónea. Aunque las posibilidades de corrección parecen ser varias,

Patrón	Tratamiento
hecha a {inf}	[-h]
	[-he][+e]
	[-hec][+ec]
	[-hech][+ech]
	[-hecha][+echa]
	[-hecha a][+echa a]

hay problemas ocultos en algunas de estas soluciones que las convierten en inviables. La primera solución realizará sustituciones indeseadas en la estructura errónea cuando el infinitivo contenga una *h-* al eliminarla incorrectamente, generando una nueva estructura errónea. Las subsiguientes soluciones tienen el mismo problema si existen infinitivos con las secuencias de caracteres *h-*, *he*, *hec-*, *hech-* y *hecha.*

Estructura errónea	Tratamiento	Estructura corregida erróneamente
hecha a **he**rvir	[-he][+e]	echa a ervir
hecha a **hech**izar	[-hech][+ech]	echa a echizar
hecha a co**hecha**r	[-hecha][+echa]	echa a coechar

Habida cuenta de estas circunstancias de incertidumbre, la última opción es la que ofrecerá, con más garantías, la solución correcta al patrón planteado.

Se deduce, por lo tanto, que es necesario establecer en el tratamiento una secuencia de búsqueda lo suficientemente extensa como para evitar coincidencias con otras partes de la estructura errónea. Este hecho obliga a optar por soluciones conceptuales en lugar de la sustitución exclusiva de los caracteres afectados para evitar posibles falsos positivos y correcciones indebidas.

Un recurso que debe considerarse en el diseño del tratamiento del error es el uso de los ESPACIOS EN BLANCO. En general, los espacios en blanco ayudan a delimitar una secuencia de búsqueda con el fin de evitar problemas ocultos. En el ejemplo anterior, hubiera sido suficiente plantear como tratamiento la palabra *hecha* y el espacio en blanco que le sigue. Este espacio en blanco permite acotar la extracción

en la cadena de búsqueda y ayuda a definir al automatismo cuál es la secuencia inequívoca sobre la que debe actuar.

Patrón	Tratamiento
hecha a {inf}	[-hecha][+echa]

La extensión del patrón y del error que se presente en él también tiene relevancia para el desarrollo de las soluciones. Cuanto más corta sea la secuencia de búsqueda mayor es el riesgo de obviar problemas ocultos. En el siguiente patrón es necesario incluir dos espacios en blanco colindantes —a izquierda y derecha— a la secuencia afectada para asegurar la ausencia de problemas ocultos. No existen formas del verbo *dar*, ni adjetivos, ni formas verbales escondidas bajo algunas de las codificaciones #dar, @{adj} o #{vb} que coincidan con la secuencia *de* que puedan dar lugar a sustituciones indeseadas.

Patrón	Tratamiento
#dar @{adj}? alegría de #{vb}	[- de][+]

Si fuera necesario se podrían incluir literales del cotexto para aumentar la fiabilidad.

Patrón	Tratamiento
#dar @{adj}? alegría de #{vb}	[-alegría de][+ alegría]

7.1.10.4 *Códigos en la solución*

Existen algunos casos en los que el tratamiento se basa en sustituciones completas del patrón por la secuencia correcta.

Patrón	Tratamiento
solamente	solamente

Sin embargo, en este subgrupo de patrones los hay que admiten flexión o conjugación en su casuística como, por ejemplo:

Patrón	Tratamiento
construído	construido
esponzorizar	patrocinar

Una forma de abordar este problema sería incluir como patrones todas las formas flexionadas afectadas para el primer patrón y todas las formas conjugadas para el segundo y establecer su correspondiente tratamiento a cada uno.

Otra forma más fiable que ahuyenta la omisión de casos por parte del lingüista que codifica es marcar el patrón y el tratamiento para que tanto el patrón como la solución cubran toda la casuística.

Patrón	Tratamiento
@construído	@construido
#esponzorizar	#patrocinar

En el primer caso, la codificación con la arroba incluye además de género y número, las flexiones apreciativas las cuales no están afectadas por este fenómeno de tildación incorrecta dado que la sílaba tónica recae en el morfema —*construidico*, *construiducho*, etc.—, por lo que siendo cuatro las formas realmente afectadas —se excluyen las apreciativas—, se opta por desarrollar cada una en un patrón independiente, en lugar de codificar con la arroba.

Patrón	Tratamiento
construído	construido
construída	construida
construídos	construidos
construídas	construidas

En el caso verbal, se propone mantener un único patrón que represente a toda la conjugación del verbo para evitar reescribirla manualmente. En este caso será necesario que el lexificador interprete adecuadamente el tratamiento propuesto para generar automáticamente la correcta relación entre los dos verbos —a cada forma conjugada del patrón le corresponde la misma forma conjugada del verbo propuesto—.

A lo largo de este trabajo se presentarán las soluciones que ofrecen más certidumbre basadas en distintos tipos de tratamiento.

Patrón	Tratamiento
`@{det}@@{masc} @{adj}? @cochambre`	`[-@@{masc}][+@@{fem}]`
`¿@{art} qué`	`[-@{art}]`
`@{ser}@@{3ª pl} suficiente con`	`[-@@{3ª pl}][+@@{3ª sg}]`
`en contra @nuestro`	`[-@nuestro][+de nosotros]`

7.1.10.5 *Generación de formas erróneas*

Por último, deben hacerse algunas consideraciones en relación con la generación de formas incorrectas inclusas en el contenido de los patrones.

En algunos casos de errores ortográficos, sin excluir a los restantes, los patrones representan a formas inexistentes, que no tienen presencia en el lexicón.

`*desvastar` `*estadunidense` `*vehiculo` `*verguenza`

Si se desea aumentar la cobertura de estos patrones a todas sus formas conjugadas o flexionadas mediante la codificación propuesta, habría que marcarlos de la siguiente forma:

`#desvastar` `@estadunidense` `@vehiculo` `@verguenza`

con ello, en teoría estarían cubiertas todas las formas conjugadas del verbo incorrecto **desvastar* y todas las formas flexionadas de los vocablos incorrectos *estadunidense*, **vehiculo* y **verguenza*.

En el caso verbal este hecho no entraña ningún problema siempre que se trata de conjugación regular debido a que existen conjugadores automáticos que son capaces de generar correctamente todas las formas conjugadas de un verbo inexistente de acuerdo con la norma ortográfica y gramatical regular; **deshechar; *desvastar, *linkear,* etc. Pero para flexionar formas a partir de una palabra ficticia, errónea, no existen aplicaciones conocidas que realicen esta tarea con fiabilidad.

Una opción automatizable que responde con fidelidad a los planteamientos de formalización expresos en esta tesis, será generar estas formas de modo automático partiendo de la información contenida en la solución. El autómata deberá tomar la forma errónea del patrón, inexistente en el lexicón, aplicarle la medida correctiva, generar automáticamente todas las flexiones o conjugaciones de la forma del patrón –ahora ya legítima y presente en el lexicón- y y sobre estas, aplicarle la solución en negativo, es decir, la ejecución deberá considerar donde haya un [+] la supresión de esa cadena, y donde haya un [–] su adición. Así partiremos de la forma canónica **vendecir* y aplicaremos la corrección hasta obtener *bendecir.* Sobre esta forma el conjugador ofrecerá todas las formas posibles de este verbo tratable y, por último, esta batería generada deberá recibir el tratamiento correctivo a la inversa, de modo que obtengamos todas las formas posibles y ajustadas a su modelo de conjugación del verbo inexistente **vendecir.*

Capítulo 8

Tipologías de error para sistemas de corrección automática

Una tipología de errores es un recurso abstracto que permite administrar y organizar toda la casuística de errores que reconoce y trata un sistema de revisión textual. La tipología, a la postre, será un instrumento que define la cobertura que ofrece el corrector. El diseño de estas taxonomías es una tarea ligada a la creación de sistemas de corrección automática y, aunque se trata de un planteamiento inicial en el desarrollo del corrector, es un recurso sometido a una continua revisión durante el proceso de desarrollo del sistema.

Uno de los hitos que deben ser alcanzados previo al proceso de dotar de contenido a PatErr o a cualquier sistema de revisión textual es el establecimiento de una tipología de errores. En el caso que aquí se expone, el repertorio de patrones codificados ha permitido afinar la tipología de los errores tratados a medida que el registro de datos incrementaba.

A pesar de los intentos de clasificación y de los proyectos desarrollados, debe reconocerse la inexistencia de una tipología de errores universal debido, en gran medida, a que diferentes lenguas dan lugar a diferentes errores que superan una clasificación universal. Por otro lado, como dice Díaz Villa:

> [...] la elaboración de una tipología de errores nunca puede ser llevada a cabo independientemente del objetivo para el cual haya sido concebida (2005:409).

Este principio que asume el desarrollo de las taxonomías *ad hoc* se convierte en otro escollo para el establecimiento de ese esquema de valor absoluto.

No obstante, a partir de las publicaciones que abordan estos asuntos, se observan ciertas tendencias y criterios que forman parte del esquema clasificatorio de varios autores y proyectos. Desde un punto de vista general y obviando la

aplicación concreta para la que se defina dicho catálogo de errores, la confección de las tipologías puede abordarse atendiendo, bien a las causas que generan esos errores, bien a las consecuencias, efectos, que tales errores producen.

Cualquiera que sea el foco elegido para el diseño de la clasificación, la tipología de cada corrector funcionará, como se sugirió, como una herramienta que permite evaluar el funcionamiento del sistema en términos de cobertura (Ramírez Bustamante *et al*., 1994; Díaz Villa, 2005).

8.1 Panorámica

Se presenta, a continuación, una breve reseña sobre las distintas orientaciones y tendencias publicadas con relación a las tipologías de los sistemas de corrección automática. Se describen, por un lado, las propuestas de algunos autores y, por otro, las soluciones que se han llevado a cabo para el desarrollo de algunos proyectos.

8.1.1 Veronis (1988)

Este autor establece una distinción entre *errores de competencia* y *errores de actuación*[84], esto es, los provocados por ciertas carencias cognitivas o los meramente fortuitos relacionados con la producción de mensajes incompletos, lagunas informativas, incorrecciones o imprecisiones que violan una gramática de la competencia. Veronis, además, establece una posible relación entre los tipos de error y los tipos de usuario —nativos y usuarios de L2— partiendo de errores relacionados con la concordancia del francés. Esta relación pareció resultar injustificada y poco fiable debido a que los usuarios, nativos o no, yerran en ambas parcelas —competencia y actuación— sin poder hacer asociaciones unívocas entre errores y usuarios (Ramírez Bustamante *et al.,* 1994).

[84] Según este autor, los errores de actuación, como los tipográficos, se deben a problemas mecánicos o neuromotores. Los asociados a la competencia, por su lado, pueden identificarse con dos hechos; desconocimiento por parte del hablante de los principios que rigen el funcionamiento general de la lengua o desconocimiento del usuario del dominio conceptual del sistema con el que está tratando (Veronis, 1988).

8.1.2 Corder (1991)

En la misma línea de Veronis, Corder trata de establecer una relación sólida entre determinados tipos de error y determinados grupos de hablantes. Previamente, distingue entre *errores* y *faltas*. Los primeros, asociados a la competencia del hablante, se deben al desconocimiento de las reglas que gobiernan el sistema de una lengua y generalmente son producidos por usuarios no nativos. Las faltas, por su parte, no implican desconocimiento lingüístico, por lo que se asocian a los usuarios nativos de una lengua. Son la resulta de lapsus de memoria, estados psicofísicos inhabituales u otras condiciones excepcionales para el hablante y están, por lo tanto, relacionados con la actuación.

8.1.3 Ramírez Bustamante *et al.* (1994)

A diferencia de Veronis, Corder y Vosse (1992), que relacionan los errores con problemas de competencia o fallos de actuación en el redactor, en Ramírez Bustamante *et al.* se desarrolla una tipología basada en los efectos lingüísticos que producen los errores, en lugar de atender a las causas que los provocan.

Este enfoque les permite abordar la categorización como la creación de un catálogo de desviaciones y violaciones que provocan igual efecto, con independencia del nivel lingüístico del que se parte y de aquel que finalmente quede afectado. En esta propuesta los errores se agrupan en (i) aquellos que violan constricciones impuestas por las reglas de la gramática del sistema, y (ii) los que violan las relaciones estructurales descritas por las reglas.

Los efectos de un error, por su parte, se describirán en términos de incompatibilidad de rasgos entre elementos o alteraciones estructurales. Estas premisas teóricas, según sus autores, *facilita*(n) *el tratamiento de cada tipo de error con diferentes técnicas de detección y corrección* (1994:580).

8.1.4 Oliva (1997)

La propuesta de este autor, además de apostar por una clasificación que escinda los errores de competencia de aquellos propios de la actuación, incluye un nuevo parámetro para controlar la cobertura del corrector; la *frecuencia* de los casos que está en disposición de tratar. Asimismo, presenta una distinción en función de la complejidad del aparato formal necesitado para la detección de cada tipo de error propuesto.

8.1.5 Verberne (2002)

A partir de un enfoque de N-gramas que desarrolla en su trabajo, establece una clasificación de errores divididos en dos grupos; por un lado, los términos no presentes en el repertorio de un idioma, *non words*, y por otro, palabras de una lengua empleadas incorrectamente en un contexto dado, *real words.* Dentro de estas últimas, Verberne propone una clasificación más exhaustiva;

- Error tipográfico.
- Asociado a los errores de actuación; *cana-vana*.
- Lapso fonético o cognitivo.
- Generalmente involucra a palabras homófonas; *ajito -agito.*
- Error gramatical.
- Anomalía semántica.
- Inserción o eliminación de palabras.
- Espacio innecesario.

8.1.6 Díaz Villa (2005)

Esta autora mantiene vigente la distinción entre errores de motivación cognitiva —competencia— y errores fortuitos derivados del uso, —actuación—. Para el caso de los primeros, sugiere el desarrollo de una clasificación pormenorizada atendiendo a las causas que generan estos errores. Díaz Villa recoge, así, cinco

tipos principales de errores gramaticales; de puntuación, de concordancia, léxicos, sintácticos y semánticos.

En el caso de los errores fortuitos, con menor relevancia para el desarrollo del corrector automático, señala esta autora que, aunque no parecen fáciles de diagnosticar —tampoco son fácilmente predecibles—, sí son, por lo general, solucionables de manera algorítmica (Díaz Villa, 2005:410).

8.1.7 Wedbjer Rambell (1999-2000)

Para esta autora las tipologías son sistemas de clasificación organizados jerárquicamente capaces de recoger todos los casos de error encontrados en una lengua (2000:5). El criterio que establece para el diseño de su propia tipología es la proximidad del contexto lingüístico o cotexto; *The basic division between error types is based on how much context is needed for an error to be recognised* (2000:5). A diferencia de los diseños examinados, el de Wedbjer Rambell presenta un sistema más exhaustivo, estructurado con mayor grado de detalle, que lo convierte en una taxonomía con un amplio grado de granularidad.

> There are five groups: spelling errors, grammar problems, punctuation problems, graphical problems, and style, meaning, and reference problems. Each group contains a number of categories which in turn are divided in subcategories (2000:5).

8.1.8 Proyecto *GramCheck* (1996)

GramCheck, basado en técnicas de alto nivel, distingue inicialmente entre errores gramaticales y deficiencias de estilo, siendo la única diferencia funcional entre ambos el hecho de que para los primeros se proporciona una corrección junto con información relacionada con el problema lingüístico en cuestión (1996:31).

No obstante, como reconocen los autores Ramírez Bustamante y Sánchez León, el estudio del corpus tratado reveló la existencia de una clasificación subyacente que ya habían propuesto en Ramírez Bustamante *et al.* (1996); (i) errores que violan restricciones estructurales y (ii) errores que violan

restricciones no estructurales. De esta doble clasificación, emerge una tipología mixta:

Errores gramaticales

- *No estructurales*: concordancia intrasintagmática e intersintagmática.
- *Estructurales*: relaciones entre núcleos y argumentos, subcategorización de preposiciones, sujeto en construcciones impersonales, usos del gerundio en sustitución de una subordinada de relativo, omisión, adición o sustitución de caracteres o signos de puntuación.

Deficiencias de estilo

- *No estructurales*: deficiencias léxicas de diversa índole; mal uso de palabras extranjeras, falsas derivaciones, mal uso de expresiones latinas, abuso de construcciones pasivas, o el uso desviado del gerundio.
- *Estructurales*: relacionadas con calcos sintácticos de otras lenguas.

8.1.9 *CON-TEXT* (1998)

CON-TEXT, basado en herramientas de análisis de bajo nivel, es capaz de detectar dos bloques de errores; los relacionados con la puntuación y los que abordan la cobertura de errores morfosintácticos (1998:168):

- *Errores de puntuación*: la herramienta diseñada para la captación de estos errores no proporciona ayuda sobre la colocación de todos los signos de puntuación, pero sí trata los signos mal ubicados que aparecen en el texto.
- *Errores morfosintácticos*: incluye reglas de concordancia que formulan secuencias de descripciones morfosintácticas erróneas con respecto al género y número dentro de sintagmas nominales. En este mismo apartado, se tratan algunas secuencias erróneas sobre formas flexivas que presentan pares casi homógrafos y que constituyen habitualmente errores cognitivos o de simple descuido.

8.1.10 *Stilus* (2002)

En Villena *et al*. (2002) se expone sucintamente el diseño y cobertura de este programa de revisión lingüística que, en buena medida, aprovecha los recursos de *CON-TEXT* (1998). Su tipología de errores gramaticales se articula, en primera instancia, en torno a los niveles lingüísticos que intervienen en las secuencias señaladas como error; nivel de la palabra, nivel morfológico, nivel sintáctico y nivel semántico. Dentro del módulo sintáctico explicitan los siguientes conjuntos de errores:

- *Errores relacionados con la concordancia,* que contravienen restricciones de las categorías gramaticales de género y número.
- *Errores de secuencias*, errores que violan restricciones de secuencialización de las categorías léxicas.

Este último grupo se compone de variados subtipos cuyo factor común es que generan secuencias ilegales. Errores relacionados con el dequeísmo, la homofonía, la vacilación de preposiciones, etc., quedarían recogidos en este subgrupo dentro del módulo sintáctico.

Como se desprende de esta breve panorámica es común, dentro de la tradición tipológica y con independencia de la premisa inicial adoptada, agrupar los errores atendiendo al nivel de análisis que se ve afectado en cada caso.

Sin embargo, como se reconoce en Ramírez Bustamante *et al.* (2000), el nivel de descripción lingüística al que pertenece un error, no siempre se corresponde con el nivel de procesamiento en que puede ser tratado, esto es, el nivel de descripción afectado no es el nivel de intervención. Es por esto por lo que, como se defiende en esta propuesta, numerosos errores relacionados con cuestiones sintácticas que se producen en contextos locales pueden ser detectados sin necesitar un análisis de alto nivel de abstracción o un procesamiento sintáctico.

8.2 Delimitación del concepto *error*

En este contexto, y a partir de estas consideraciones, puede definirse el error como una cadena desviada con respecto de otras bien formadas. Por lo tanto, y retomando la propuesta de Ramírez Bustamante *et al.* (1996, 1998) un segmento presentará un error cuando refleje (i) una desviación estructural con respecto de cualquiera de las estructuras positivas descritas por la gramática del corrector o (ii) una violación de una constricción previamente consignada en la gramática del sistema.

Pero al margen de esta definición, que se fundamenta en criterios estructurales, puede definirse el error en otros términos que limiten su alcance conceptual. En el marco de este trabajo, cuando se hace referencia al concepto de *error* se hará en un sentido amplio, con el ánimo de abarcar tanto las desviaciones estructurales, SECUENCIAS AGRAMATICALES, como las INCORRECCIONES, entendidas estas como secuencias que deben evitarse según las orientaciones de la norma.

Aunque, en rigor, conceptos como *gramaticalidad*, *propiedad* o *adecuación* deban presentarse como fenómenos distinguidos en ámbitos como el de la sintaxis,

> La gramaticalidad de las expresiones representa una propiedad constitutiva e interna relativa a su naturaleza formal, mientras que la corrección responde a factores regulativos de carácter social (Bosque y Gutiérrez, 2009:32).

para los fines computacionales que rigen este trabajo se ha optado por subsumir todos estos fenómenos como incidencias que tendrán asociadas una etiqueta de error.

A partir de este planteamiento, los errores de PatErr podrán asociarse, bien a cuestiones estructurales, bien a asuntos relacionados con el uso. Se abordarán, por lo tanto, casos en los que se vea comprometida la estructura de una oración, agramaticalidad, y casos en los que se detecten deficiencias en el modo en que una secuencia se asocia a otra o se incrusta en el discurso. Puede relacionarse con este ámbito de la propiedad o la adecuación, el término *incorrección*, que siguiendo la pauta académica se aplicará a *secuencias atestiguadas que deben evitarse en el uso culto* (DPD: *agramaticalidad*).

Cualquiera que sea el caso, el tratamiento que proveerá PatErr se ajusta a las consideraciones y orientaciones propuestas por las fuentes académicas, ya sea en el terreno gramatical, ya en el estilístico.

8.3 Propuesta de tipología de errores

Otro de los parámetros que se manejan para la arquitectura de las tipologías de error que operan en este tipo de programas es el GRADO DE SEVERIDAD DEL ERROR detectado (Díaz Villa, 2005).

En nuestro caso, y separándonos de la tradición teórica que se intuye de los otros proyectos examinados, es este el criterio que vertebrará transversalmente la clasificación, y dará lugar a los tres bloques de incidencias que ya han sido mencionados; errores, recomendaciones y avisos lingüísticos.

El contenido de estos tres bloques, esto es, los desvíos que la máquina es capaz de detectar, se organizarán, además, según los niveles lingüísticos involucrados en la incidencia. Este será el otro principio estructurador de los datos que dan cuerpo y lógica al contenido de PatErr.

Por otro lado, dentro de los criterios de diseño que deben considerarse en la creación y organización de este tipo de clasificaciones es el grado de GRANULARIDAD. Para evitar que la tipología resulte demasiado burda y poco funcional, es conveniente generar subclasificaciones que permitan, por un lado, manejar la información en bloques organizados y, por otro, tener una referencia más minuciosa de la cobertura del sistema.

Una tipología que recoja una cantidad de datos significativa requiere, para su mejor manejo y explotación, que se organice garantizando un amplio grado de granularidad. Este concepto parte del principio de que es más fácil reutilizar unidades —de información, en nuestro caso— más pequeñas abriendo la posibilidad de seleccionar o descartar aquellos rasgos informativos que sean de interés. Asumiendo este principio esencial en el diseño de programas y bases de datos, y siguiendo las orientaciones de Wedbjer Rambell (1999-2000), en este trabajo se propone una suerte de taxonomía vertical que remite a los niveles

lingüísticos afectados por el error, y que es atravesada por la clasificación transversal que emerge del grado de severidad del error.

Como se puede advertir, la escisión principal que se plantea en cuanto a la organización fenomenológica de los errores no se establece a partir de las causas —error de competencia o actuación—, ni de los efectos causados por las desviaciones o violaciones que presenta el material "corregible" —violación de restricciones estructurales o no estructurales—, ni tampoco de la cantidad de cotexto necesario para que el algoritmo lleve a cabo la corrección. Los parámetros que servirán de andamio para la tipología de errores serán, por un lado, el grado de severidad del error, de donde surgen tres grandes núcleos de actuación —*error, recomendación* y *aviso*— y por otro, el nivel lingüístico en el que repercute la interferencia; *ortografía, tipografía, gramática, léxico, morfología* y *estilo*.

Unificando estos dos criterios, podemos bosquejar un esquema con 18 componentes:

	ERROR	**RECOMENDACIÓN**	**AVISO**
ORTOGRAFÍA	error de ortografía	recomendación de ortografía	aviso de ortografía
TIPOGRAFÍA	error de tipografía	recomendación de tipografía	aviso de tipografía
GRAMÁTICA-SINTAXIS	error de gramática-sintaxis	recomendación de gramática-sintaxis	aviso de gramática-sintaxis
MORFOLOGÍA	error de morfología	recomendación de morfología	aviso de morfología
LÉXICO	error de léxico	recomendación de léxico	aviso de léxico
ESTILO	error de estilo	recomendación de estilo	aviso de estilo

8.3.1 Grado de severidad

Partiendo de la severidad que presente la incidencia identificada por el programa, la escala gradual cuenta en un extremo con la etiqueta de *error*, entendido como fenómeno que se presenta en una palabra o expresión que está vedado por la gramática o la norma del español.

Los otros dos subconjuntos de incidencias que completan este rango rebajan el grado de severidad del error y suponen una novedad en el diseño y prestaciones de este tipo de sistemas autómatas.

En lo que respecta a las *recomendaciones*, se prevén para aquellos casos en los que, aunque la lengua ofrezca varias opciones correctas, solo una es la recomendada por la norma. Será el usuario el que tome la última decisión sobre la intervención —o no— sobre la secuencia identificada por el sistema.

El *aviso lingüístico*, por último, conforma el otro extremo en la graduación de severidad de errores que aquí proponemos. Los avisos conforman un bloque de orientaciones preventivas que el usuario debe contemplar en ciertos contextos en los que habitualmente se registran desvíos, y con ciertas formas que son susceptibles de generar errores. El aviso es una herramienta que, aunque no corrige el error lo ahuyenta y asegura la re-visión sobre una secuencia que debe ser analizada por el escritor.

El repertorio de avisos lingüísticos está relacionado mayoritariamente con casos que requieren una interpretación semántica o un análisis sintáctico de alto nivel; adjunción de papeles temáticos, rección verbal con restricciones semánticas, etc. Para estos casos complejos, el sistema ofrecerá información sobre los usos correctos, reglas y restricciones que la gramática o la norma impone o propone y a los que deberá ceñirse la secuencia identificada por el programa.

Para generar la batería de recomendaciones y sugerencias que puede prestar PatErr se han tenido en cuenta las variedades lingüísticas panhispánicas, y se han observado y codificado ciertos rasgos pertinentes relacionados con criterios diafásicos, diastráticos y diatópicos. En estas sugerencias, además, se proponen alternativas a la secuencia detectada que permiten perfeccionar el discurso.

8.3.2 Niveles lingüísticos

Cada nivel lingüístico que se ha propuesto es el título de la coalición de fenómenos que se colectan de un mismo estrato de la lengua. Los niveles contienen, a su vez, otros menores que para fines de organización interna permiten una estructura de los datos más minuciosa, interconectada y manejable.

Este rasgo de diseño revierte en la granularidad del sistema. De este modo, cada caso de desvío recogido en PatErr queda perfectamente definido y separado de los demás casos, gracias a que cada entrada tiene atribuido un haz de rasgos relevantes que lo sitúan en un lugar concreto dentro del sistema. Así, un caso como el siguiente,

PATRÓN	ACCIÓN CORRECTIVA	NIVEL LINGÜÍSTICO	FENÓMENO ASOCIADO	VARIEDAD	REGISTRO
estuvistes	[-stes][+ste]	morfológico	Morfología verbal.	1	vulgar
					normal

presenta un error de morfología verbal —fenómeno asociado— provocado por la epéntesis de una *-s* —incidencia 1018—, propio del registro vulgar —registro y variedad— que afecta al plano morfológico —nivel lingüístico—. Como se puede intuir, el registro de una descripción minuciosa de cada caso permite tener un sistema más robusto, enriquecido, manipulable y útil tanto para la investigación como para la acción de revisión textual.

Por último, debe señalarse que para simplificar el elenco de niveles susceptibles de corrección o revisión se han subsumido en un bloque dos niveles de la lengua; el léxico y el semántico bajo el membrete de *léxico*. Los contenidos recogidos por esta etiqueta estarán escindidos, —granulados—, a nivel interno dentro de los campos de incidencia y fenómeno asociado.

A continuación, se ofrece una taxonomía aproximada de los fenómenos que se abordan en el actual estado de desarrollo de PatErr. No pretende ser sino ilustrativa, por lo que hay casos tratados que se escapan a esta clasificación.

ORTOGRAFÍA

1. Homófonos
    a. Pares habituales
    b. Tratamientos complejos: *a-ha; e-he; iba-iva; hecho-echo; deshecho-desecho; ahí-ay-hay*
2. Casi homófonos
3. Ortografía de algunas palabras susceptibles de generar error
4. Palabras que admiten dos grafías

MORFOLOGÍA

1. Prefijación
2. Sufijación
3. Flexión de número
    a. Sin flexión de plural
    b. Pluralia tantum
    c. Singularia tantum
    d. Plural de sustantivos especiales
    e. Plural de monosílabos
    f. Plural de expresiones compuestas
    g. Plural de latinismos y helenismos
4. Flexión de género
5. Algunos superlativos
6. Errores en la flexión nominal
7. Composición
    a. «Vb+N»
8. Morfología verbal
    a. Epéntesis
    b. Error en construcción de una forma verbal
    c. Vacilación entre tiempos o modos verbales

GRAMÁTICA

1. Concordancia
    a. Concordancia gramatical intrasintagmática: categorías
    b. Concordancia gramatical intersintagmático: relaciones entre sujeto y verbo
    c. Concordancia de género
    d. Concordancia de número
    e. Concordancia de persona
2. Gramática verbal
    a. Usos del gerundio
    b. Usos del infinitivo
    c. Vacilación entre tiempos o modos verbales
    d. Usos no impersonales de verbos impersonales
3. Dequeísmo
4. Queísmo
5. Supresión indebida de preposiciones
6. Errores de régimen preposicional en términos o expresiones
7. Errores de régimen preposicional en verbos
8. Errores de régimen preposicional en verbos que rechazan preposición.

LÉXICO

1. Impropiedades léxicas
2. Restricciones semánticas
3. Precisión
4. Segmentación y unificación

   a. Sin cambio de significado
   b. Con cambio de significado: *sino-si no*; *por qué- porque- por que*
5. Interferencias entre construcciones
6. Extranjerismos
7. Locuciones latinas
8. Neologismos
9. Coloquialismos
10. Vulgarismos

TIPOGRAFÍA

1. Doble opción en una expresión
2. Doble opción en palabras
   a. Superlativos
3. Construcción de expresiones
   a. Expresiones partitivas
   b. Expresiones recíprocas
   c. Expresiones reflexivas
   d. Expresiones numerales
   e. Expresiones de enfoque
4. Expresiones de tiempo: adverbios y preposiciones
5. Cacofonías
6. Coloquialismos
7. Vulgarismos
8. Redundancia

TIPOGRAFÍA

1. Extranjerismos
2. Expresiones y términos latinos
3. Pautas de uso de las mayúsculas
4. Pautas de puntuación
5. Segmentación: el guion
6. Abreviaturas
7. Siglas

Capítulo 9

PatErr: esquema de anotación

El desarrollo del recurso que ofrece esta propuesta necesita un andamiaje que aporte estructura y orden a la masa de datos que conforman todo su contenido. PatErr, entendido como un repositorio de errores codificados, se alojará según las especificaciones que se fijarán a continuación, en una base de datos que será accesible y manipulable tanto para el lingüista, como para el programador. El esquema de anotación será, por lo tanto, ese constructo teórico que permita organizar eficazmente todas las dosis de información que constituyen el núcleo de este recurso.

Una de las tareas que debe acometerse inicialmente, en paralelo a la compilación de las fuentes bibliográficas y la selección de los datos que pretenden tratarse es decidir un boceto de ESQUEMA DE ANOTACIÓN que recoja la planificación del tipo de información formalizada que se pretende almacenar y la organización de los datos objeto de registro. En el esquema de anotación deben quedar explícitos los parámetros lingüísticos y atributos que deban ser registrados y formalizados junto con el patrón de error codificado y la solución que se le prevea dar. Se ha utilizado la forma *boceto*, con toda la intención, porque el esquema de anotación es, como la tipología de errores, un diseño que se va perfilando a medida que la casuística crece y el desarrollo del recurso va revelando nuevas necesidades y exigencias.

En el esquema de anotación se deberán tener previstos ciertos aspectos que afectan tanto al contenido como a la forma del recurso. Toda tipología de errores destinada a ser tratada computacionalmente conlleva indefectiblemente, varias clasificaciones subsidiarias que permiten varios niveles organizativos resultantes de los distintos enfoques. Cuantas más clasificaciones se hagan, mayor será la granularidad del sistema que contiene los datos y, por lo tanto, mayor será la accesibilidad a los datos para su tratamiento o manipulación. Se obtienen, así, subgrupos representados por niveles de descripción lingüística, por variedades o

registros lingüísticos, por grupos temáticos de incidencia, etc. según el enfoque y filtro que se aplique a la base de datos. Cada una de estas clasificaciones, —campos en la base de datos—, deberá quedar perfectamente acotada y ceñida, además de definir el tipo de información que va a recoger y representar. Será necesario, pues, establecer el tipo de etiquetas, clasificaciones y taxonomías que pongan orden en el contenido de PatErr.

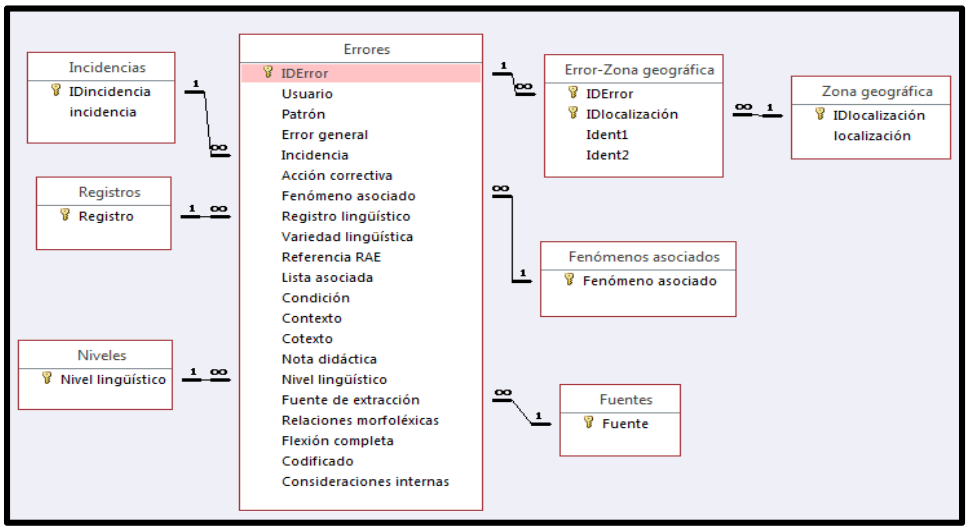
Junto con estas cuestiones formales relacionadas con la organización y clasificación de datos, en el esquema de anotación deberán quedar explícitos criterios lingüísticos como el número y límites de las categorías gramaticales con las que se va a trabajar. Para la propuesta que se ofrece en este trabajo se adoptará el etiquetario que se ha propuesto para el Lexicón TIP.

9.1 Alojamiento

El espacio que recoge toda la información que constituye el repositorio de errores es una base de datos relacional. Ha sido necesario, por lo tanto, determinar claramente las tablas y campos que la componen, definirlos en forma adecuada con un nombre y acotar el tipo de información que van a contener y pueden recibir.

De este modo, se facilita la búsqueda de datos por filtros que permiten crear subconjuntos que en algún punto del proceso pueden recibir un tratamiento similar. Los registros funcionan, pues, como un esquema o andamio sobre el que se organizan los datos que componen toda la información relativa a un mismo patrón de error.

La siguiente figura muestra las relaciones que se han establecido entre la tabla que contiene los errores y otras tablas subsidiarias que se han asociado a ella.



9.2 Estructura

Para el desarrollo de este recurso con plenas garantías, el lingüista que asuma la tarea del desarrollo de patrones deberá conocer todas las especificaciones y formatos relacionados con la información que debe registrarse en la base de datos. Estos criterios, que se concretan en el esquema de anotación, deberán ser explícitos y detallados para mantener la consistencia y la posibilidad de manipulación de los datos.

El esquema que se ha diseñado para el registro de información de la tabla principal de PatErr que hemos denominado *Errores* presentará los siguientes campos.

9.2.1 Error

Este campo está compuesto por valores numéricos que representan a cada patrón de error. Son valores únicos, irrepetibles e irrepetidos. Hasta el momento, la base de datos cuenta con algo más de 2900 casos registrados. Este número, no obstante, no representa la cantidad de errores que se pueden tratar, sino la cantidad de patrones —con sus consiguientes registros— que hay formalizados y codificados.

Como se ha visto, un mismo patrón, puede servir como base de error y corrección de cientos de errores; sirva como ejemplo un patrón como el siguiente;

`os #{vb}##{2ª pl} ustedes`	`[-os #{vb}##{2ª pers pl}]` `[+se #{vb}##{3ª pers pl}]`
**Os calláis ustedes y no se hable más.*	

Esta entrada, gracias al sistema generativo de casuística activado por el lenguaje de codificación, ofrecerá cobertura a cualquier secuencia que contenga la expresión errónea «*os* —cualquier verbo en 2º persona del plural— *ustedes*»; **Os largáis ustedes que son los que están/estáis sobrando*.

9.2.2 Patrón

En este campo se alojan los patrones de error. Está compuesto por palabras, secuencias de palabras o estructuras codificadas que contienen un error. Cada nueva entrada en PATRÓN genera un nuevo registro de datos con 18 campos asociados que contienen información relativa a ese patrón concreto.

1865	@{det}@@{fem} tangas
1888	@{pn}@@{átono} #{vb}##{imperativo}
1890	#marchar a
1891	#marchar para
1892	#pudrir
1894	yo y @{pn}@@{sujeto}
1895	yo y @{det}@@{pos}+ @{N}
1908	ustedes @{adv}? os #{vb}##{2ª pl}
1909	os #{vb}##{2ª pl} ustedes

El contenido que recoge este campo puede representarse exclusivamente a sí mismo o puede dar lugar, a partir de la codificación, a una batería de patrones relacionados con esa misma estructura.

Como ha quedado expreso, el número de patrones no se corresponde ni con el número de errores que PatErr puede procesar, ni con el número de fenómenos que quedan tratados y solucionados. Un fenómeno como el dequeísmo, por ejemplo, cuenta con más de 50 entradas que codifican toda la casuística relacionada observando los datos que nos ofrece el análisis lingüístico. De este se deduce que no es un fenómeno exclusivamente verbal toda vez que puede surgir en expresiones como **a menos de que, *una vez de que, *primero de que,* etc. Por otro lado, en el contexto verbal, hay ciertas consideraciones que deben observarse a nivel computacional; verbos como *avergonzarse, lamentarse, alegrarse, sorprenderse,* etc., tienen un régimen especial como se intuye en oposiciones presentes entre la primera y la tercera persona del singular: *Me avergüenzo de que vengas* ~ **Me avergüenza de que vengas*. Por último, debe tenerse en cuenta que entre la forma verbal y *de que* puede intercalarse algún adverbio, **Me avergüenza muchísimo de que.*

Con esta batería de algo más de 50 patrones registrados se le da cobertura a cientos de errores reales, si pensamos que cada verbo con el que puede surgir, presenta 221 variantes; 77 formas simples, 71 formas simples con pronombres enclíticos y 73 formas compuestas, y el lexicón registra más de 71 000 adverbios Las combinaciones posibles —posibles errores— que el programa puede identificar se multiplican exponencialmente, aunque todos están asociados a un mismo fenómeno, el dequeísmo.

9.2.3 Error general

Este campo recoge etiquetas que expresan, declarativamente, el error sin marcas de codificación, esto es, se formula el error de forma simplificada para facilitar su lectura e identificación en la base de datos. Esta información es útil para la tarea del lingüista que codifica porque le permite crear subconjuntos dentro de la masa de datos que comparten rasgos temáticos o reciben un tratamiento similar.

IDE	U	PATRÓN	ACCIÓN CORRECTIVA	ERROR GRAL	NIVEL LINGÜÍSTICO	FENÓMENO ASOCIADO
1656	ly	sores	[-es]	pl de palabras invariables	morfológico	Morfología: construcción de plurales.
1657	ly	dones @{N}@@{propio}	[-es]	pl de palabras invariables	morfológico	Morfología: construcción de plurales.
1658	ly	sanes	[-es]	pl de palabras invariables	morfológico	Morfología: construcción de plurales.
1659	ly	frayes	[-es]	pl de palabras invariables	morfológico	Morfología: construcción de plurales.
1660	ly	frays	[-s]	pl de palabras invariables	morfológico	Morfología: construcción de plurales.

Como cabe esperar, el campo ERROR GENERAL es un repertorio abierto que se puede incrementar con la suma de nuevos patrones. No obstante, como se ve en la tabla precedente, el número de errores generales diferentes es menor que el de patrones.

9.2.4 Acción correctiva

En este campo se desarrolla la corrección o la opción recomendada —en los casos de recomendación o aviso— que se debe llevar a cabo. Este espacio contiene la información ejecutiva que el programa ha de llevar a cabo para desarrollar con éxito la corrección y devolverle al usuario una opción satisfactoria. Esta información se expresa codificada y se basa en el procedimiento básico de la sustitución de los segmentos afectados por otros que subsanan el error.

El campo de ACCIÓN CORRECTIVA es también un campo abierto que se incrementa a medida que crece el número de patrones y se amplía la cobertura de PatErr. No obstante, no es extraño, como se observa en la imagen que se ofrece, que patrones que tratan un mismo fenómeno, aun teniendo estructuras diferentes, puedan recibir la misma instrucción para llevar a cabo la corrección.

PATRÓN	ACCIÓN CORRECTIVA	FENÓMENO ASOCIADO
spray	[-spray][+esprái]	Extranjerismos: palabra o expresión
bajo @{det}+ @base de	[-bajo][+sobre]	Construcción de una expresión.
hay @{adv}? que @{vb}@@{pn reflex}@@{2ª pl}	[-#{vb}@@{pn refl}@@{2ª pl}][+#{vb}@@{pn refl}@@{3ª sg}]	Gramática verbal: impersonalidad.
hay @{adv}? que @{vb}@@{pn reflex}@@{2ª sg}	[-#vb}@@{pn refl}@@{2ª sg}][+#{vb}@@{pn refl}@@{3ª sg}]	Gramática verbal: impersonalidad.
hay @{adv}? que @{vb}@@{pn reflex}@@{1ª pl}	[-#{vb}@@{pn refl}@@{1ª pl}][+#{vb}@@{pn refl}@@{3ª sg}]	Gramática verbal: impersonalidad.
hay @{adv}? que #{vb}@@{pn reflex}@@{1ª sg}	[-#{vb}@@{pn refl}@@{1ª sg}][+#{vb}@@{pn refl}@@{3ª sg}]	Gramática verbal: impersonalidad.
hubiereis ^({participio})	[-hubiereis][+fuereis]	Gramática verbal: impersonalidad.
hubiéremos ^({participio})	[-hubiéremos][+fuéremos]	Gramática verbal: impersonalidad.
hubisteis ^({participio})	[-hubisteis][+fuisteis]	Gramática verbal: impersonalidad.
hubimos ^({participio})	[-hubimos][+fuimos]	Gramática verbal: impersonalidad.
hubieseis ^({participio})	[-hubieseis][+fueseis]	Gramática verbal: impersonalidad.
hubiésemos ^({participio})	[-hubiésemos][+fuésemos]	Gramática verbal: impersonalidad.

Por otro lado, habrá casos, generalmente asociados a los avisos lingüísticos, en los que este registro quedará en blanco por no haber una acción correctiva certera aplicable de forma automática. Para solventar estos patrones es necesaria una interpretación activa del texto y será el usuario, a partir de la información proporcionada por la nota didáctica, el que lleve a cabo la corrección o los retoques que considere pertinentes.

PATRÓN	NOTA DIDÁCTICA
sendos	Aunque es común su utilización en el lenguaje popular de América como sinónimo de "enorme, descomunal", este uso es rechazado por la norma cult
sendas	Su significado es "uno para cada una de las personas o cosas mencionadas", y por su propia significación, solo se utiliza en plural. No debe emplearse
sendos	Su significado es "uno para cada una de las personas o cosas mencionadas", y por su propia significación, solo se utiliza en plural. No debe emplearse
en especies	Para hacer referencia a la expresión "ganar un sueldo a cambio de género, no de dinero" la forma correcta es "en especie"; <i>Cuando vivía en el Con
a dosis	Cuando esta expresión se utiliza como complemento circunstancial de verbos como: administrar, suministrar, ingerir, tomar... o similares, la forma cor
del tirón	La forma correcta para referirnos a la expresión que indica "de una vez, de un golpe" es "de un tirón".
@cuanto más	El parecido fónico entre "cuando más" y "cuanto más", hace que en numerosas ocasiones haya interferencias entre estas expresiones; así cuando má
cuando más	El parecido fónico entre "cuando más" y "cuanto más", hace que en numerosas ocasiones haya interferencias entre estas expresiones; así cuando má

9.2.5 Nivel lingüístico

Esta etiqueta permite vincular a cada error con el nivel lingüístico en el que se inscribe. La taxonomía que se ha desarrollado para este recurso es la siguiente;

Gramatical-sintaxis: incidencias relativas a las normas de estructura de constituyentes y viabilidad —gramaticalidad— de la frase. Actualmente hay 760 patrones codificados y etiquetados en este nivel.

Léxico: patrones relacionados con el significado y sentido de las palabras o expresiones; abarca también incidencias relativas a la semántica. Cuenta con 318 patrones codificados etiquetados y adscritos a este nivel.

Morfológico: incidencias relacionadas con las reglas de composición de las palabras. Actualmente hay 234 patrones asociados a este nivel.

Ortográfico: errores o desvíos relacionados con las reglas y convenciones que rigen la escritura de nuestra lengua. En este punto de desarrollo, hay 1 208 patrones codificados de errores ortográficos cuya comisión incide sobre otros niveles de la lengua.

Tipográfico: fenómenos relacionados, bien con el conjunto de reglas de estilo y escritura tipográfica, bien con aspectos relacionados con los signos de puntuación. Actualmente hay 113 patrones incluidos en este nivel.

Estilo: incidencias relacionadas con cuestiones de estilo como la pertinencia, adecuación, elegancia y/o la precisión de ciertas palabras o expresiones presentes en un texto escrito sujeto a la norma culta del español. Actualmente hay 254 patrones codificados relacionados con cuestiones estilísticas.

NIVEL LINGÜÍSTICO es un campo cerrado que solo ofrece la posibilidad de inscripción del patrón de error en una de las seis opciones previstas en la taxonomía.

9.2.6 Incidencia

Este término hace referencia al fenómeno lingüístico del que participa un error o un uso de la lengua ajeno a la norma. Cada caso registrado, cada patrón de error, lleva asignada una incidencia enunciada mediante un valor numérico que localiza a ese patrón dentro del elenco de posibles incidencias. Este valor numérico se asocia a un membrete que recoge el *título* del fenómeno objeto del tratamiento. Este número, además de organizar la masa de datos, indica, por su valor, qué tipo de incidencia se ha detectado; error (1-1999), recomendación (2001-3999) y aviso (4001-5999). Los rangos numéricos previstos pueden contener hasta dos mil entradas por cada bloque de incidencias. Su organización es la que sigue:

9.2.6.1 *Errores*

- 1-999: Este rango recoge errores concretos. La actuación en estos casos se lleva a cabo sobre un término o fenómeno específico. Se recogen en este bloque casos como los errores que generan los homófonos *echo* y *hecho*.
- 1001-1999: Registro de los errores generalizados, de aplicación masiva que inciden sobre varios elementos erróneos. Un ejemplo de estos dentro del ámbito de la gramática verbal pueden ser los usos erróneos del gerundio. Estos usos afectan de igual manera a todos los verbos en su forma de gerundio, por lo que tratamiento de este error puede generalizarse y expandirse.

9.2.6.2 *Recomendaciones*

- 2001-2999: Recomendaciones concretas sobre un patrón determinado. En estos casos la corrección que se lleva a cabo no es automática sino interactiva, en el sentido de que será el usuario el que tome la decisión final sobre el texto. No obstante, una recomendación de estilo a propósito de la expresión desaconsejada por la norma culta a día de hoy puede ser automatizada para que sustituya esta expresión por alguna de las opciones sugeridas; actualmente, hoy en día, etc.
- 3001-3999: Esta horquilla recoge los casos en los que una recomendación puede operar a un nivel más general y abstracto. Un ejemplo de estas recomendaciones masivas es el tratamiento tipográfico de los extranjerismos.

9.2.6.3 *Avisos lingüísticos*

- 4001-4999: Avisos concretos, que funcionan del mismo modo en que lo hacen las recomendaciones individualizadas. En la mayoría de los casos no podrán ofrecer sino orientaciones de carácter general que ayuden al escritor a subsanar el error tras una interpretación activa del texto. Un caso de estos avisos es el patrón dedicado al distributivo *sendos*, que informa sobre la extensión y restricciones semánticas que impone esta forma. Como puede advertirse, el contenido de estos no es automatizable.
- 5001-5999: Estos avisos, de aplicación generalizada a todos los casos que queden afectados por el patrón registrado, sigue con la lógica de esta taxonomía. Se recogen en este rango casos relacionados con la morfología verbal derivados de la vacilación entre formas o construcciones verbales.

En el Anexo 10 se ofrece el contenido desgranado de la tabla de INCIDENCIAS. Como se podrá observar, hay varios fenómenos que aparecen codificados en los tres grupos, —errores, recomendaciones y avisos—, debido a que, dependiendo del cotexto en el que se inscriba el fenómeno, el tratamiento será diferente.

9.2.7 Fenómeno asociado

Este campo contiene un listado con el título de todos los fenómenos lingüísticos que son tratados por el programa; son pues, etiquetas más genéricas que las propias de INCIDENCIA. El registro de este membrete global permite la organización de toda la casuística codificada por repertorios temáticos, ya sean errores, recomendaciones o avisos. Resulta muy útil disponer de esta información organizada y explicita para poder extraer por bloques los temas tratados y tener un control sobre la cobertura que se ofrece de cada fenómeno; error de concordancia, dequeísmo, régimen preposicional, imprecisiones semánticas, construcción de expresiones, etc.

Como cabe esperar, es un campo abierto capaz de recibir sin límite las nuevas incorporaciones que surgen a medida que la cobertura de PatErr se acrecienta. En el Anexo 11 puede consultarse la tabla que contiene todos estos registros de FENÓMENO ASOCIADO.

9.2.8 Registro lingüístico

El campo REGISTRO LINGÜÍSTICO recoge las variedades lingüísticas en las que se pueden inscribir los patrones de error. Estos registros están asociados a una determinada situación comunicativa, a un determinado nivel estilístico o a un subcódigo relativo a un lenguaje especial.

La gama de registros puede llegar a ser muy amplia y las taxonomías fluctúan en número de elementos y criterios para recoger los diversos registros que un mismo hablante emplea según dos parámetros esenciales; la situación comunicativa y la relación que mantiene con su interlocutor.

Para los fines de este programa, solo se recogerán algunas variedades que se presentan a nivel diastrático —código restringido y elaborado— y a nivel diafásico —coloquial, oral, publicitario, etc.—. El repertorio de variantes diatópicas quedará registrado en la tabla VARIEDAD LINGÜÍSTICA. Las opciones que han sido contempladas para el programa son ocho;

- científico-técnico
- coloquial
- culto
- neutro
- oral
- periodístico
- publicitario
- vulgar

No obstante, este es un campo abierto que permite incrementar o afinar el repertorio de registros con más etiquetas que podrán ser mostradas al usuario.

9.2.9 Variedad lingüística

El campo VARIEDAD recoge información relativa a las variedades diatópicas del español. Contiene una lista de zonas geográficas donde pueden localizarse ciertos errores, usos o variantes que se asociarán a los patrones de error.

En el Anexo 12 pueden consultarse las regiones hispanohablantes previstas en este campo. Como sucede con el repertorio de REGISTRO, este también puede ser ampliado, bien por exigencias —inobservadas— de la casuística, bien por motivos de máxima precisión[85]. Esta información relativa a la variedad diatópica habrá de servir en el futuro como criterio para establecer *perfiles de usuario*. De este modo el escritor, con anterioridad a la revisión textual, podrá diseñar un perfil propio que el sistema deberá contemplar en el curso de la corrección con el fin de ajustar las opciones posibles al texto concreto.

9.2.10 Contexto

Este campo recoge información relevante sobre los contextos en los que puede y suele surgir el error. El contenido lo constituyen especificaciones que aportan

[85] Etiquetas como la de *Zona noroccidental de España* podrá granularse, por ejemplo, en registros individuales como *León, Galicia* y *Asturias*.

datos más concretos que los ofrecidos para la variedad diafásica en la tabla de REGISTRO LINGÜÍSTICO.

Generalmente, este campo aportará información sobre el ámbito o sublenguaje en el que típicamente podrán identificarse esos patrones de error; informático, deportivo, legal, medios de comunicación, etc. Por otro lado, ofrecerán claves sobre el contexto semántico en el que suele localizarse el error; partes del cuerpo, expresiones de pasado, contexto numeral, etc.

PATRÓN	CONTEXTO	NIVEL LINGÜÍSTICO
En segundo lugar, señalar que	típicamente periodístico oral	estilo
En primer lugar, señalar que	típicamente periodístico oral	estilo
@dolor al {Lista 33}	parte del cuerpo	gramatical-sintaxis
@dolor a @{det}+ {Lista 33}	parte del cuerpo	gramatical-sintaxis
@rejuvenecedor	publicidad	léxico-semántico
@antiedad	publicidad	léxico-semántico
#barajar	literal y metafórico	léxico-semántico

A diferencia de los campos anteriores, cuyo contenido era de obligado cumplimiento para cada nuevo patrón de error que ingresa en la base de datos, la información relativa al contexto comunicativo solo se hace explícita cuando se considera relevante y útil, bien para el usuario en el caso de los sublenguajes, bien para el lingüista o el programados en casos en los que son necesarias dosis de información semántica para acotar el ámbito de búsqueda de ese patrón.

9.2.11 Cotexto

Este campo recoge datos referidos al entorno lingüístico que rodea al patrón de error y excede las especificaciones presentes en el patrón. Esto sucede en casos en los que es necesario aludir a elementos presentes en la secuencia susceptible de contener un error que no se presentan en un contexto de contigüidad. COTEXTO recoge información interna orientada a que el programa, en el proceso de rastreo, pueda *cotextualizar* el patrón para extraerlo y tratarlo adecuadamente.

El cotexto actúa como un filtro que fortalece la precisión en la identificación de patrones en aquellos casos en los que la ambigüedad —en alguna de sus posibilidades— pueda estar presente. Esta información es relevante tanto para el

informático, que debe considerar formalmente el entorno del patrón, como para el lingüista que se sirve del cotexto para establecer matices en el rastreo que garantizarán la eficacia y la precisión en el proceso de identificación del error.

9.2.12 Condición

Aplicando la misma lógica que en el caso que precede, el campo CONDICIÓN ofrece información relevante —para explotación interna— que ha de considerarse para llevar a cabo los procesos de ejecución previstos por el sistema de revisión que define PatErr. Se trata de premisas, indicaciones y restricciones, que el programador debe contemplar en algunos patrones para lograr que el tratamiento se desarrolle correctamente sobre las formas y contextos precisos.

INCIDENCIA	PATRÓN	CONDICIÓN
47	@cuánto menos	en correlación con menos
47	@cuánto peor	en correlación con peor
34	hasta atrás	en textos no de México
34	hasta adelante	en textos no de México
553	#darse @{adv}+? cuenta que#	entre la negación y el vb debe haber pn; es un verbo pr
522	#arroyar	excepto 1 pers sg; arroyo
522	#cayar	excepto formas amb (cayado, cayo,...)
2003	@Tejas	excepto tejano
516	#como que sí.	fin de la frase, seguido de punto

9.2.13 Lista asociada

Este campo hace referencia a una lista de elementos sobre los que debe aplicarse el mismo tratamiento que plantea el patrón codificado. Aunque, como se verá en los capítulos dedicados al contenido del PatErr, el número de la lista está presente y con marcas de codificación en el campo patrón, el registro de este campo facilita al lingüista la extracción y control de datos.

Esta tabla, como otras que atañen al preprocesamiento del texto —contexto, cotexto, condición, etc.—, va aumentando su número de registros a medida que van surgiendo patrones de error que permiten la aplicación de este sistema globalizador de la casuística.

9.2.14 Relaciones morfoléxicas

El campo de RELACIONES MORFOLÉXICAS ofrece únicamente dos alternativas en el proceso de codificación; *sí/no.* Esta información debe ser interpretada por el programa para llevar a cabo una expansión de la casuística derivada de la palabra afectada y recogida en el patrón de error. Esta expansión permite ampliar el tratamiento previsto no solo a la palabra presente en el patrón, sino también a todas sus derivadas compuestas y parasintéticas que deban asumir el criterio de la acción correctiva. Para la propuesta que aquí se presenta, se ha contado con los resultados del trabajo de Carreras Riudavets esbozado en Santana Suárez *et al.,* (2004).

PATRÓN	ACCIÓN CORRECTIVA	RELACIONI	CONDICIÓN
@expléndido	[-expléndid][+espléndid]	Sí	
Méjico	[-j][+x]	Sí	
Oajaca	[-j][+x]	Sí	
Tejas	[-j][+x]	Sí	excepto tejano
Hispano América	[-Hispano América][+Hispanoamérica]	Sí	
Hispano-América	[-Hispano-América][+Hispanoamérica]	Sí	
Ibero América	[-Ibero América][+Iberoamérica]	Sí	
Ibero-América	[-Ibero-América][+Iberoamérica]	Sí	
Latino América	[-Latino América][+Latinoamérica]	Sí	
Latino-América	[-Latino-América][+Latinoamérica]	Sí	
Sud América	[-Sud América][+Sudamérica]	Sí	
Sud-América	[-Sud-América][+Sudamérica]	Sí	
Sur América	[-Sur América][+Sudamérica]	Sí	
Sur-América	[-Sur-América][+Sudamérica]	Sí	

La marca positiva asociada a este campo en la base de datos permite que, en casos como *Méjico*, para el que la norma recomienda la grafía *México* se lleve a cabo una corrección expansiva a todas las palabras detectadas en el texto que se relacionen morfoléxicamente con ella; *mejicano, mejicana, mejicanismo, mejicanistas,* etc.

9.2.15 Referencia normativa

En este espacio se recoge la información relativa a la fuente normativa de la RAE que avala el tratamiento e información que aporta PatErr para cada caso

identificado. Esta etiqueta se mostrará al usuario para que pueda consultar, contrastar o ahondar en el asunto objeto de la corrección, recomendación o aviso.

PATRÓN	REFERENCIA RAE	FENÓMENO ASOCIADO
#{vb} el por qué	NGLE 43.3p	Construcción de una expresión.
#{vb} el por qué	NGLE 43.3p	Construcción de una expresión.
Sino #{vb}	OLE 561	Léxico-Semántico: palabras que se
no solo \w+, si no	OLE 561	Léxico-Semántico: palabras que se
ni \w+ si no	OLE 561	Léxico-Semántico: palabras que se
no \w+ si no	OLE 561	Léxico-Semántico: palabras que se
@{det}@@{masc} @{adj}? si no	OLE 561	Léxico-Semántico: palabras que se
dado @{det}@@{masc pl}+ @{N}@@{masc pl}	DPD dar	Concordancia gramatical: suj-vb.

Las fuentes normativas en formato tradicional que han sido compiladas en este registro son las que siguen:

- *Gramática descriptiva de la lengua española*. (1999); GDLE.
- *Diccionario panhispánico de dudas*. (2005); DPD.
- *Nueva gramática de la lengua española*. (2009); NGLE.
- *Nueva gramática de la lengua española*: *Manual*. (2010); Manual.
- *Ortografía de la lengua española*. (2010); OLE.

Las fuentes consultadas en su versión en línea son:

- *Diccionario de la lengua española.* (23ª ed.). Consultado en http://dle.rae.es/?id= [de 2014-2017].
- *Diccionario panhispánico de dudas*. Consultado en http://lema.rae.es/dpd/?key= [de 2014-2017].

Este repertorio podrá incrementarse con otras obras de referencia, especialmente en el caso de los errores relacionados con la ortotipografía[86].

9.2.16 Fuente de extracción

Con esta etiqueta se identifica la fuente de la que se ha extraído el caso o fenómeno erróneo que será objeto de tratamiento. Estas fuentes, como se ha expuesto, pueden ser tanto referencias bibliográficas, como el Corpus TIP.

[86] Son de máximo rigor y vigencia, en los temas relacionados con cuestiones ortotipográficas, los tratados de Martínez de Sousa (2000, 2004).

Las posibles etiquetas que hasta el momento se han registrado en este campo son;

- 500 —Las 500 dudas más frecuentes del español—
- Blog —Blog de Lengua—
- Corpus (TIP)
- DPD
- DRAE
- LEC —Libro del español correcto—
- Manual
- NGLE
- OLE
- otros

Como en otros casos, este campo puede aumentar sus elementos a medida que se incremente el volumen de patrones de error.

PATRÓN	FUENTE DE EXTRACCIÓN
uno de ambos	NGLE
cualquiera de ambos	NGLE
más @mayor que	500
@fuertísimo	NGLE
cluquillas	DPD
quizás	NGLE
inclusives	NGLE
@{det}@@{masc sg}+ @{adj}? taxis	NGLE
@expléndido	OLE

9.2.17 Nota didáctica

El campo dedicado a las notas y comentarios lingüísticos recoge el repertorio de glosas didácticas que han sido elaboradas par ser asociadas a cada patrón de error. Estas notas se han previsto para que el usuario disponga de la información de uso y ejemplos pertinentes para abordar el proceso interactivo de revisión y corrección o, en última instancia, para aumentar su pericia lingüística en el uso del español.

Estas glosas, junto con la referencia de la fuente normativa, forman la batería de recursos didácticos necesarios para llevar a cabo el servicio de asesoría que el PatErr puede ofrecer al escritor. Así, un patrón identificado que trate algunos de los aspectos de la segmentación de las formas *si no/sino* devolverá al usuario la siguiente nota:

> Cuando *sino* introduce un elemento que reemplaza al negado en la oración precedente, la grafía correcta es la compacta: *No vi toda la exposición sino lo más importante*.
>
> El valor de esta conjunción adversativa puede variar según el enunciado; además de contraponer una idea afirmativa a otra negativa expresada anteriormente, en algunos casos denota adición enfática, mientras que en otros puede sustituirse por *más que, salvo* o *excepto*. No debe confundirse esta conjunción adversativa *sino*, con la secuencia de conjunción condicional y negación *si no*, donde la negación es tónica, a diferencia del [*no*] átono de *sino*: *No sé si no tendrá que dimitir*.

Por otro lado, las notas que se han registrado para esta propuesta son neutras o no marcadas, en el sentido de que están pensadas para cualquier escritor del español actual y no están orientadas a un usuario con un perfil concreto, como pueda ser el de un estudiante de primaria, de secundaria, de ELE o un traductor. No obstante, y como trabajo futuro, se puede contemplar la posibilidad de reajustar y adaptar estas notas a diferentes perfiles que presenten circunstancias o necesidades específicas.

9.2.18 Consideraciones internas

Este campo está dedicado a notas y observaciones asistemáticas de carácter interno que son de importancia para los desarrolladores, tanto lingüistas como informáticos. El contenido que presenta es variado, y surge de las exigencias o problemas que se manifiestan en el proceso de registro, codificación y estudio de la casuística. Generalmente contendrá excepciones, problemas generados por la ambigüedad o notas que sirven para acotar o informar sobre algunos fenómenos que se han detectado en el proceso de desarrollo de la base de datos.

El diseño y estructura informativa que se ha planteado para la creación y constitución de PatErr lo convierten en una herramienta flexible y en continuo desarrollo. El programa que se constituya a partir de él o se nutra de la información que recoge podrá determinar perfiles tipo de usuario, o bien permitir una configuración *ad hoc* por parte de este. Así el hablante nativo, el aprendiz de español, el estudiante nativo de lengua española, el profesional de la lengua —traductores y correctores—, los usuarios de español americano, etc. podrán seleccionar el tipo de corrección, recomendaciones y avisos que deseen recibir, así como los niveles lingüísticos que quiere que sean revisados en su texto.

9.3 Metodología derivada

Una vez establecido y desarrollado el esquema de anotación que recoja los parámetros lingüísticos y atributos que deban ser codificados y el modo de codificación —qué formato y lenguaje común se desarrollará para la comunicación entre lingüistas e informáticos para la manipulación de datos—, se emprenderá el proceso de catalogación y codificación de todos los fenómenos que se pretenden tratar. Deberá etiquetarse cada fenómeno, —junto con los que deriven de él y sus posibles flexiones— con la identificación del nivel lingüístico al que afecta, la variedad de español a la que pertenece —tanto en el eje sincrónico como en el diacrónico—, el registro que le es propio en el caso del léxico, el fenómeno lingüístico implicado —dequeísmo, neologismo, concordancia, restricciones semánticas, etc.— así como otras consideraciones relevantes en términos de análisis lingüístico. Asimismo, será necesaria la redacción de una breve nota explicativa que aporte al usuario la información gramatical útil relacionada con la incidencia detectada.

La puesta en marcha de esta metodología tendrá como consecuencia la constitución de un arsenal de errores codificados tratables por la máquina que, como ya se ha advertido, podrá ser la base o el complemento de un programa de revisión textual.

Capítulo 10

PatErr: glosario de los errores tratados

A partir de estas páginas se ofrece un glosario que recoge algunos de los errores que constituyen el contenido y cobertura de PatErr. Se ha comprobado que una buena parte de los asuntos que aquí se plantean no encuentran tratamiento en otros correctores que se han contrastado[87].

Se ha optado por ilustrar los casos tratados ofreciendo el resultado práctico —en forma de patrones de error formalizados y codificados— que derivan de la metodología y planteamientos que se han expuesto en capítulos anteriores.

La presentación se organizará en capítulos que se corresponden con los niveles lingüísticos afectados por el error que se aborda. El orden de exposición obedece al orden en el que deben aplicarse los patrones sobre el texto. Como puede intuirse, como norma general sujeta a excepciones, el primer tratamiento debe llevarse a cabo sobre el nivel ortográfico y posteriormente el morfológico. Aplicados estos patrones, se podrán tratar errores los errores gramaticales, léxicos y estilísticos que involucren a palabras bien formadas y reconocidas por el lexicón.

Habida cuenta del amplio grado de granularidad que presenta el diseño y la tipología de errores de aquí proponemos, el hecho de que haya solapamientos y subsunciones en algunos fenómenos o de que exista la posibilidad de adscribir un mismo patrón a diferentes incidencias es una coyuntura habitual.

Por este motivo, para la exposición de la casuística compilada se ha optado por no segmentar o fragmentar los fenómenos o asuntos relacionados, motivo por el que se ha dedicado un capítulo a algunos temas transversales que aglutinan el

[87] Además del contraste de los resultados con el corrector de Word 2016, se han llevado a cabo cotejos con otros programas disponibles en la red:
http://spanishchecker.com/
http://www.mystilus.com/Correccion_interactiva
https://languagetool.org/es/

contenido de algunos fenómenos que se han abordado cuyo tratamiento exige la intervención en más de un nivel de la lengua. Esta unificación fenomenológica ha provocado que la exposición de algunas cuestiones relacionadas con la tipografía haya quedado diseminada a lo largo del tratamiento de otros asuntos, en lugar de ocupar un capítulo propio.

Por motivos obvios de espacio y tiempo no se ha podido reescribir todo el contenido alojado en la base de datos que contiene cerca de 3000 registros con 18 campos con información que deriva del trabajo de ingeniería lingüística que se ha llevado a cabo. Pero la profusión de información que se ha recopilado acerca de la amplia cantidad de fenómenos que se han tratado impide llevar la exhaustividad descriptiva y expositiva —no la investigativa, ni la ejecutiva— al extremo.

En cada capítulo y epígrafe se ha hecho énfasis en los aspectos que en la tarea de desarrollo han resultado más escurridizos, reseñables, problemáticos, anómalos o incluso curiosos. Se aportan ejemplos que encarnan lo abstracto de muchos patrones y se ha combinará la explicación gramatical con la formalización en términos de patrón de error[88].

La batería de ejemplos que se presenta, como cabe esperar, está compuesta por secuencias erróneas, ya sea por motivos *gramaticales*, ya sea por otros fenómenos relacionados con la *corrección.* Los primeros se marcarán conforme a las convenciones tradicionales con el símbolo [*]. Los segundos, que representarán estructuras o formas incorrectas y/o desaconsejadas por la norma estarán señalizados por la marca [⊗].

Se ofrecen, además, las medidas correctoras previstas, la casuística posible del error, las glosas lingüísticas que se ofrecerán al usuario final, etc. El criterio de exposición ha sido dictado por el foco de interés que presenta cada dato. Por este motivo, en algunos casos se expondrá el patrón o la batería de patrones codificados, mientras que en otros en los que el patrón es fácilmente deducible y no precisa matiz se expondrán solo las listas de elementos que se han codificado.

[88] Las descripciones gramaticales que se presentarán para acompañar a la exposición de los patrones de error se han tomado de las fuentes normativas (NLGE, *Manual,* DPD). No se harán explícitas las referencias en cada caso concreto que se trata, aunque esta información está registrada en la base de datos asociada a cada error. Sí se señalarán, por el contrario, aquellos argumentos o teorías tomadas de fuentes distintas de las académicas.

Uno de los objetivos que motiva esta exposición es servir como punto de partida o guía —en ningún caso definitiva— para el desarrollo de programas que asuman la perspectiva que aquí se ha adoptado; la revisión textual a partir de la identificación de patrones de error formalizados y codificados.

A lo largo de las páginas que siguen se pretende, además, acercar la materia lingüística, las palabras, al entorno y realidad de la máquina. Y es en ese acercamiento donde, recurrentemente, aparecen temas como la ambigüedad de las formas y los sintagmas o la carencia de interpretación semántica, que obligan a adoptar nuevas perspectivas desnaturalizadas en términos lingüísticos con el fin de encontrar soluciones programables a problemas lingüísticos en este nuevo entorno computacional.

Capítulo 11

Ortografía

Aunque una explotación óptima del recurso que presentamos requiera un tratamiento ortográfico previo a su aplicación, son muchos los casos en los que se observan errores ortográficos de amplia frecuencia que suelen pasar desapercibidos para otros sistemas correctores.

Esta notable ausencia puede deberse a que, en ocasiones, los errores se materializan en formas que constituyen elementos legítimos del lexicón de partida. Este hecho imposibilita que sin un tratamiento específico para estas formas, que indefectiblemente partirá del recurso del cotexto, un corrector ortográfico pueda detectar estas incidencias.

A continuación, se abordarán errores ortográficos relacionados con términos homófonos, casi homófonos y otros que admiten dos grafías, aunque solo una es la elegida y recomendada por la norma. Con respecto al primer grupo de asuntos, se propone un tratamiento específico para pares ya clásicos como *hech*o y *echo* y sus derivados *desecho* y *desecho*, *ahí- hay- ay* o *e-he* y *a-h*a. Estos planteamientos y metodología servirán de modelo para intuir cómo se han desarrollado otros tratamientos que, por motivos de espacio, no podrán formar parte de este trabajo —*haya-halla*; *vaya-valla*—.

11.1Homófonos

Este epígrafe recogerá casos de homófonos —no homógrafos— que, por su confluencia acústica en el plano oral, con frecuencia salpican los textos de errores ortográficos. Pares habituales en el mundo de la corrección como *hojear-ojear* o *revelar-rebelar,* junto con las variantes derivadas de las formas *hecho* y *desecho*,

hay y *ahí*, etc., habrán de ser tratados en este bloque. Aunque estas vacilaciones derivan, a la postre, en errores ortográficos son innegables las implicaciones que la homofonía tiene en el plano léxico-semántico.

El mayor problema que se plantea para ofrecer un tratamiento correcto a estas formas es que prácticamente todas las opciones con la que se trabaja son legítimas como palabras del español, esto es, todas son entradas de cualquier lexicón. Este hecho producido por la ambigüedad se convierte en el mayor escollo que debe sortear tanto el programa como el programador.

Ante estas limitaciones, y tras un estudio exhaustivo de los cotextos que suelen asociarse a estas formas homófonas, se corregirán solo aquellos casos en los que el tratamiento automático se efectúe con todas las garantías, sin propiciar falsos positivos, ni intervenir en secuencias correctas. Para el resto de los casos que no pueden ser corregidos de un modo automático, se han diseñado avisos lingüísticos que asesorarán al escritor para que sea él, de forma interactiva, el que lleve a cabo la corrección.

A continuación, se desarrollarán las medidas que se han previsto para estos casos de homofonía en dos bloques diferenciados. En primer lugar, se acometerán las formas y contenido de los avisos lingüísticos. Seguidamente, se expondrán los fenómenos y patrones de error sobre los que se podrá llevar a cabo una intervención automática en el texto.

11.1.1 Avisos lingüísticos: algunos pares habituales

Se aportan, a continuación, los pares de formas homófonas junto con el contenido de sus glosas que han sido objeto de escrutinio. Ante la detección por parte del programa de alguna de estas formas, se activará el aviso lingüístico que proporcionará la información léxica y de uso necesaria para que el usuario, atendiendo a su intención comunicativa, tome las medidas correctivas convenientes.

@acerbo	Este adjetivo hace referencia a algo *áspero al gusto*. Su homófono, *acervo,* es un sustantivo que significa *conjunto de bienes morales o culturales acumulados por la tradición*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.
@acervo	Este sustantivo significa *conjunto de bienes morales o culturales acumulados por la tradición*. Su homófono, *acerbo*, es un adjetivo que hace referencia a algo *áspero al gusto*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.

#bacilar	Este adjetivo hace referencia a un tipo de bacterias. Su homófono, *vacilo*, es la 1ª persona del singular del presente del verbo *vacilar* que significa por un lado, *titubear, estar indeciso* y, por otro, *engañar, tomar el pelo, burlarse o reírse de alguien*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.
#vacilar	Este verbo significa por un lado, *titubear, estar indeciso* y, por otro, *engañar, tomar el pelo, burlarse o reírse de alguien*. Su homófono, *bacilar*, es un adjetivo que hace referencia a un tipo de bacterias, los bacilos. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.

bacilo	Este sustantivo es el nombre de un tipo de bacteria. Su homófono, *vacilo*, es la 1ª persona del singular del presente del verbo *vacilar* que significa, por un lado *titubear, estar indeciso* y, por otro, *engañar, tomar el pelo, burlarse o reírse de alguien*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.
vacilo	Esta forma es la 1ª persona del singular del presente del verbo *vacilar* que significa, por un lado *titubear, estar indeciso* y, por otro, *engañar, tomar el pelo, burlarse o reírse de alguien*. Su homófono, *vacilo*, es un sustantivo que hace referencia a un tipo de bacterias. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.

@basto	Este se utiliza para hacer referencia a algo *grosero, tosco, sin pulimento*. Su homófono *vasto*, significa *dilatado, muy extenso o muy grande*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.
@vasto	Este adjetivo se utiliza para aludir a algo *dilatado, muy extenso o muy grande*. Su homófono *basto*, significa a *grosero, tosco, sin pulimento*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.

#hojear	Este verbo, que deriva del sustantivo *hoja* significa *pasar las hojas de un libro o cuaderno leyendo deprisa algunos pasajes*. Para hacer referencia a *mirar a alguna parte, mirar superficialmente un texto*, el término correcto es *ojear*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.
#ojear	Este verbo, que deriva de *ojo*, significa *mirar a alguna parte, mirar*

	superficialmente un texto. Para hacer referencia a *pasar las hojas de un libro o cuaderno leyendo deprisa algunos pasajes*, el término correcto es *hojear*, que deriva de *hoja*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.

iba	Se trata de una forma verbal correspondiente a la 1ª y 3ª persona del singular del imperfecto de indicativo del verbo *ir*. Su homófono, *iva* o *IVA*, es el acrónimo de *impuesto de valor añadido*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.
iva	Este acrónimo (o su variante escrita en mayúsculas, IVA) se refiere al *impuesto de valor añadido*. Su homófono, *iba*, es una forma verbal correspondiente a la 1ª y 3ª persona del singular del imperfecto de indicativo del verbo *ir*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.

#rebelar	Este verbo significa *sublevar, levantar a alguien haciendo que falte a la obediencia debida*. Su homófono *revelar*, tiene por significado *descubrir o manifestar lo ignorado o secreto.* Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.
#revelar	Este verbo significa *descubrir o manifestar lo ignorado o secreto*. Su homófono *rebelar* tiene por significado *sublevar, levantar a alguien haciendo que falte a la obediencia debida*. Aunque ambas palabras suenan igual, sus significados y grafías son diferentes.

11.1.2 Pares frecuentes con tratamiento automático

En este epígrafe quedan recogidos por bloques los patrones de error que ejecutarán, de forma automática, la corrección de las formas homófonas que causen error en el texto.

Se tratan, en primera instancia, los pares confusos *a-ha* y *e-he,* y los homófonos *iva-iba*. Para su tratamiento, se han desarrollado patrones que recogen algunos de los contextos posibles (e identificables) de los que pueden participar estas formas.

Expuestas estas incidencias, se ofrecerá, de forma individualizada, el tratamiento que se ha previsto para los errores derivados de las siguientes formas:

- *echo*, *hecho*
- *desecho*, *deshecho*
- *hay, ahí, ay*

11.1.2.1 *a, ha*

Para solventar algunas de las múltiples incidencias que se pueden generar a partir del uso escrito tanto de la preposición *a* como de la forma verbal *ha* se han codificado las siguientes entradas:

a {participio}	[-a] [+ha]	**No a llegado el esperado paquete.*
a de {inf}	[-a de] [+ha de]	**A juzgar por el bullicio a de haber mucha gente en la plaza.*

11.1.2.2 *e, he*

El par compuesto por la conjunción *e* y la forma conjugada *he* se resuelve con los siguientes patrones:

e {participio}	[-e][+he]	**¿Cuándo e dicho yo semejante barbaridad?*
e de {inf}	[-e][+he]	**Según mi médico e de hacer ejercicio con más frecuencia.*

11.1.2.3 *iba, iva*

Otro de los pares que con frecuencia se presentan como un foco irradiador de errores son los generados por el sustantivo acrónimo y la forma verbal *iba*.

@{det}@@{masc sg}+ @{adj}? iba	[- iba][+ iva]
**El iba cultural ha sido inasumible para muchos bolsillos.*	

al iba	[- iba][+ iva]
**Al iba hay que sumarle otras tasas ineludibles.*	

del iba	[- iba][+ iva]
**El impuesto del iba es una herramienta para los políticos.*	

@{pn}@@{reflex}@@{1ª y 3ª sg} iva	[- iva][+ iba]
*Se iva sin pagar.	

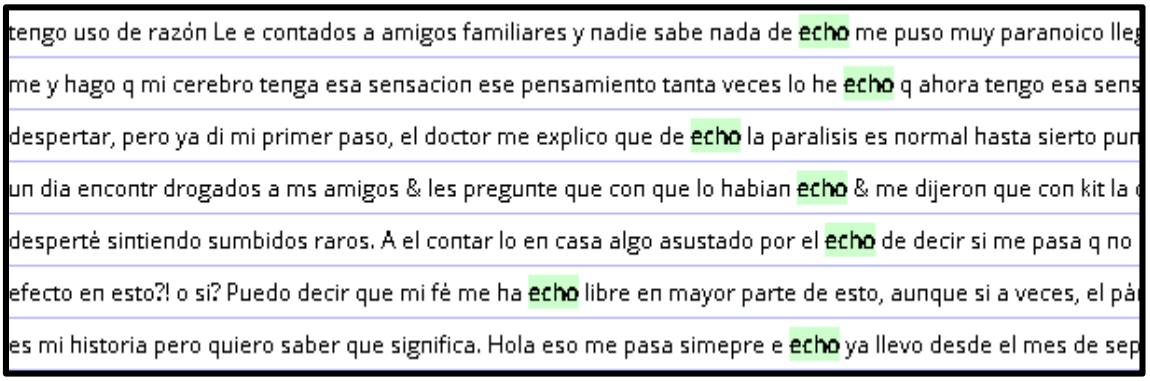
@{pn}@@{sujeto}@@{1ª y 3ª sg} iva	[- iva][+ iba]
*Me iva al Caribe de buena gana.	

Estos patrones son complementarios al aviso lingüístico del epígrafe anterior que se activará siempre que estos patrones de error no hayan podido ejecutarse previamente.

11.2 *hecho* y *echo*

Es frecuente confundir el sustantivo masculino *hecho,* con el presente de indicativo del verbo *echar, echo* y el participio del verbo *hacer, hecho*. Las formas verbales, además, presentan homófonos en algunas de sus versiones flexionadas; *hecha-echa* y *hechas-echas.*

Se expone, a continuación, una muestra extraída del corpus de frases que sirve para ilustrar la profusión y confusión que genera, en este caso, la forma *echo.*


tengo uso de razón Le e contados a amigos familiares y nadie sabe nada de echo me puso muy paranoico lleg
me y hago q mi cerebro tenga esa sensacion ese pensamiento tanta veces lo he echo q ahora tengo esa sens
despertar, pero ya di mi primer paso, el doctor me explico que de echo la paralisis es normal hasta sierto pun
un dia encontr drogados a ms amigos & les pregunte que con que lo habian echo & me dijeron que con kit la
desperté sintiendo sumbidos raros. A el contar lo en casa algo asustado por el echo de decir si me pasa q no
efecto en esto?! o si? Puedo decir que mi fé me ha echo libre en mayor parte de esto, aunque si a veces, el pá
es mi historia pero quiero saber que significa. Hola eso me pasa simepre e echo ya llevo desde el mes de sep

Para desenmarañar este panorama de un modo automático y ágil, y sin contar con información semántica, ha sido necesario codificar un amplio grupo de patrones; 141 de error y ocho avisos lingüísticos.

A continuación, se expone una topografía de los contextos que, conforme a las reglas de la gramática, pueden ocupar estas formas.

11.2.1 Contextos de las formas participiales: *hecho, hecha, hechos, hechas*

El participio del verbo *hacer* en su forma masculina y singular es requerido para formar tanto los tiempos verbales compuestos *ha hecho, hemos hecho,* etc. como las perífrasis de participio del tipo *estar hecho un*…, *ir hecho un*… En estas últimas configuraciones también participan las formas femeninas y plurales del participio en concordancia de género y número con el sujeto; *Mariví vino hecha un cuadro, Estos muchachos están hechos unos hombres.*

11.2.2 Contextos del sustantivo *hecho*

Como cualquier sustantivo puede ser núcleo de un grupo nominal, por lo que generalmente irá precedido de un determinante que lo habilite como núcleo; *Todos aquellos hechos demostraban su buena voluntad; Aquel hecho inesperado marcó mi vida*.

11.2.3 Contextos de la locución *de hecho*

Esta locución, que generalmente se hace seguir de una coma, se hace imprevisible en cuanto a sus contextos debido a que se puede combinar con cualquier categoría gramatical. No obstante, la preposición servirá como anzuelo para la identificación del patrón de error en el texto.

11.2.4 Contextos de las flexiones de *echar*: *echo, echa* y *echas*

Estas formas corresponden a la primera —*echo*—, segunda —*echas*— y tercera —*echa*— persona del singular del presente de indicativo del verbo *echar* y a las

correspondientes formas pronominales de *echarse* —exentas del pronombre clítico—.

El verbo *echar*, además, forma parte de locuciones de uso muy frecuente como: *echar de menos*, *echar una mano, echar a suertes*, etc.

Habida cuenta de este panorama, se han propuesto varias medidas correctivas a partir del diseño de patrones específicos que individualmente subsanan parte del universo de errores derivado del uso escrito de estas formas homófonas.

El proceso de codificación se ha basado en la captura parcial de errores siguiendo un orden de certeza y yendo de los casos más simples de detectar a los más complejos. Estos últimos requieren, cuando menos, estar acotados por un cotexto preciso que evite falsos positivos. De este modo se irá resolviendo este asunto de homofonía progresivamente, hasta ir agotando toda la casuística documentada.

Junto a la complejidad que estos fenómenos representan para el PLN, convive la insistente acción de la ambigüedad que en algunos casos y contextos muy concretos no permite la automatización de la corrección. Secuencias como *echa la salsa; hecha la salsa*, solo podrán ser procesadas a partir de una interpretación semántica del texto y de la situación comunicativa. En estos casos en los que la ambigüedad se multiplica, el programa solo ofrecerá un aviso lingüístico que proporciona al escritor las claves necesarias para garantizar la elección correcta de la forma pertinente.

Partiendo de estas circunstancias, se ha dividido el contenido de estos patrones en dos bloques. Se ofrecen, en primer lugar, los patrones que subsanan los errores relacionados con el uso del participio del verbo *hacer*. En segundo lugar, se presentan los patrones que detectan los errores generados en expresiones compuestas por formas del verbo *echar.*

11.2.5 Errores que involucran formas participiales de *hacer*

Para solventar los errores causados por el primer bloque de formas —participios del verbo *hacer*— la primera medida que se ha considerado es la inclusión de un patrón de error que trate la forma **echos.*

echos	hechos

Esta forma, inexistente en español, será errónea allá donde se presente y con independencia del sentido que el escritor tenga previsto concederle. Tanto en secuencias en las que funcione como un sustantivo, *hechos consumados*, como cuando sea un participio de plural, *Llegaron hechos polvo*, la forma **echos* quedará automáticamente corregida por su forma correcta *hechos*.

Otra regla de cotexto laxo que soluciona múltiples errores y debe aplicarse en un primer rastreo del programa es la siguiente:

#haber {adv}? echo	[- echo][+ hecho]

En esta expresión compleja compuesta por el verbo *haber,* la forma siempre será la propia del participio de *hacer; hecho.*

A partir de este par de patrones de validez indubitada, los errores y sus contextos empiezan a resultar escurridizos para la máquina; hay muchos contornos sintácticos idénticos en los que pueden aparecer tanto el verbo *hacer* en su forma participial, como el verbo *echar* en su forma finita de primera persona; *echa la cuenta; hecha la suma*.

Tras un estudio de la casuística y de la viabilidad de aplicación de patrones de error, se ha concluido que si la secuencia no va precedida de formas como *haber*, *estar*, *ir,* etc. será imposible —contando solo con datos sintácticos y sin una interpretación semántica de alto nivel— saber si esa secuencia contiene un error.

Pares legítimos como los que se ofrecen a continuación pretenden dar una muestra ilustrativa de estas limitaciones insuperables para el programa[89]:

hacer/echar + SN (det + N)

con el pelo hecho un asco ⇔ *con el dosificador echo unas gotas*
hecha la sopa ⇔ *echa la sopa*

hacer/echar + SP (prep + SN/N)

hecha a mano ⇔ *echa a Federico*
hecho de verdad ⇔ *echo de menos*

hacer/echar + pn

hecho esto ⇔ *echa mucho*
hecho todo ⇔ *echo todo*

hacer/echar + adv

hecha concienzudamente ⇔ *echa siempre en falta*

Esta realidad lingüística obliga, para conseguir cierto éxito en el tratamiento de estas formas, a codificar las expresiones y locuciones más frecuentes en español que puedan generar error por la homofonía entre el participio de *hacer* y las formas finitas *echo, echa* y *echas* del verbo *echar*.

El siguiente patrón básico sirve de base para otras perífrasis similares.

`#ir echa`	`[-echa][+hecha]`
`#ir echas`	`[-echas][+hechas]`
`#ir echo`[90]	`[-echo][+hecho]`

[89] Esta imposibilidad declarada para detectar los errores en los que se involucran estas formas homófonas surge tras un estudio de la naturaleza de los sustantivos, adverbios y preposiciones que siguen a estas formas. Tras la investigación se ha concluido que no se encuentra ninguna pauta generalizable en las unidades que siguen a estas formas.

[90] Estos patrones no incluyen la forma **echos* porque ese error queda subsanado con el patrón genérico que trata exclusivamente esta forma.

Otras expresiones recogidas se recogen en la siguiente tabla. Se agrupan entre paréntesis las tres formas —*echo, echa* y *echas*—, aunque en la base de datos están recogidas individualmente.

`#venir (echo|echa|echas)`	**Ha venido echa una furia.*
`#llegar (echo|echa|echas)`	**Llegó echo añicos.*
`#ir (echo|echa|echas)`	**Se fue echa un basilisco.*
`#quedar (echo|echa|echas)`	**Ha quedado echo polvo.*
`#dejar (echo|echa|echas)`	**Nos dejaron echas un cuadro.*
`#ser (echo|echa|echas)`	**Todo fue echo con mucho cuidado.*
`#estar (echo|echa|echas)`	**Están echas unas mujercitas.*

Además de estos patrones, análogos en cuanto a sus constituyentes, se han codificado otros que recogen expresiones erróneas formadas por el participio de *hacer* seguido de otras unidades de diversa categoría.

***echa**	realidad posible a sí misma y derecha
***echas**	realidad posible a sí mismas y derechas
***echo**	realidad posible a sí mismo y derecho

Otras expresiones de uso frecuente que pueden ser portadoras de error son las encabezadas por un adverbio:

*mal *bien *muy *totalmente *absolutamente	**echa**
*mal *bien *muy	**echas**

*totalmente *absolutamente	
*mal *bien *muy *totalmente *absolutamente	**echo**

Se ha estudiado la viabilidad de un patrón de error genérico:

@adv echo

que corrigiera la forma añadiéndole la *h*, pero se ha desestimado por encontrar casuística que presenta esta configuración y no constituye un error; *Nunca echo de menos a quien no me valora; Casi echa a su hermano de casa; Rápidamente echas a correr,* etc.

Aunque se ha llevado a cabo esta codificación individual de las expresiones más frecuentes en el español escrito se han codificado, además, avisos lingüísticos de carácter general que se activarán cuando el programa detecte alguna de las siguientes secuencias:

(echo|echa|echas) @{det}+? @{N}	Para referirse al participio del verbo *hacer* en expresiones como *hecho polvo, hecho realidad, un poema hecho canción*, la forma correcta presenta la *h-* inicial.
(echo|echa|echas) @{det}+? @{N}	Para hacer referencia a alguna forma del verbo *echar*, con el sentido de *hacer que algo vaya a parar a alguna parte, dándole impulso*, la grafía correcta rechaza la *h-*; *echar agua al vino*, *echar la siesta*, etc.

Se impone, como restricción para estos patrones de aviso, que el sustantivo que sigue a la forma no sea ambiguo en cuanto a su categoría. Se evita así la activación del aviso ante secuencias comunes y correctas como *echo de menos, hecha la maleta,* etc. Tanto *de,* —preposición y sustantivo—, como *la* —artículo y sustantivo— son partículas ambiguas de alta frecuencia que suscitarían innecesariamente la activación del aviso.

11.2.6 Errores que involucran formas flexionadas de *echar*

Estos errores derivan de las interferencias causadas por las formas participiales de *hacer*. Es habitual encontrar en la escritura las formas del verbo *echar* con *h-* inicial no solo en los homófonos que comparte con *hacer,* sino en el resto de formas flexionadas.

Siguiendo el mismo criterio en el proceso de codificación que se mantuvo en el caso anterior, se diseñan, por un lado, algunas reglas generales sin cotexto y por otro, expresiones idiomáticas —colocaciones— de uso frecuente que permiten una corrección automatizada sin necesidad de interpretaciones de alto nivel.

Para abordar el primer bloque, se codifican dos reglas generales de aplicación prioritaria cuya corrección se basa en la extracción de la *h* inicial:

`#hechar`	**Hechaba de menos aquellos días.*
`#hecharse`	**Se ha hechado novia y se ha olvidado de nosotros.*

Debe advertirse que, previa a la codificación de estos patrones, será necesaria la creación de estas conjugaciones erróneas y por tanto inexistentes en el lexicón. De este elenco, deberán excluirse las formas que coinciden con los participios *hecho, hecha, hechas*, para evitar que el programa corrija estas formas irrestrictamente cuando estén escritas conforme a la gramática.

Esta exclusión sacrifica la cobertura de la regla y deja sin tratar mucha casuística que se registra con las formas homófonas; **hecha broncas, *hecho de menos, *hecha un piropo, *hechas a perder.* Ante esta limitación se ha desarrollado, como atenuante, una batería de patrones que contienen estructuras gramaticales y expresiones idiomáticas de alta frecuencia compuestas por el verbo *echar* en sus formas finitas *echo, echa* y *echas.*

`(echo|echa|echas) a {inf}`	**Hechas a reír con mucha facilidad*
`@{pn}@@{pers} @{adv}? (echo|echa|echas)`	**Él nunca hecha la primitiva*
`@{pn}@@{reflex} (echo|echa|echas)`	**Se hecha encima todos los problemas.*

En este último patrón que contiene la etiqueta de pronombres reflexivos —elenco que incluye las formas tanto de complemento directo, como de indirecto según el diseño del etiquetario que hemos propuesto— debe extraerse la forma *lo.* Se evitan así falsos positivos en oraciones como *Lo hecho por Pepe fue heroico.*

Por último, se han registrado patrones de error con las formas *hecho, hecha* y *hechas,* que constituyen expresiones frecuentes en nuestra lengua. Quedan, de este modo, cubiertos los casos que han quedado excluidos por las restricciones aplicadas a lo largo del proceso de codificación.

echo|echa|echas @@{det}+?	**echo|echa|echas**	**echo|echa|echas**	**echo|echa|echas**
@bronca	@raíz	del trabajo	atrás
@vistazo	flores	de casa	abajo
@siesta	barriga	a suerte	
@carta	leña al fuego	de menos	
@pulso	pestes	en falta	
@cuenta	un mal de ojo	en cara	
@culpa	un capote		
@mano	un cable		
@tiento	una mano		
@poco de	un ojo		
@poquito de			

Esta tabla, incompleta aunque representativa, deberá ser nutrida con más componentes que serán revelados a partir de los datos reales que ofrece el corpus.

11.2.7 Errores que involucran al sustantivo *hecho*

Como puede intuirse este grupo de incidencias es más reducido y fácilmente tratable.

Frente a la multiplicidad de cotextos que presentan las formas verbales tratadas, este sustantivo masculino reduce los márgenes del error —quedan

excluidas las formas femeninas— y permite el desarrollo de una solución simple y eficaz desarrollada por un par de patrones:

`@{det}@@{sg}@@{masc} @{adj}? echo`	**El horrible echo acontecido nos dejó perplejos.*
`@{det}@@{pl}@@{masc} @{adj}? echos`	**No acepto tu análisis de los echos.*

Es necesario acotar en este patrón el número —masculino— del determinante para evitar que, por causas de ambigüedad categorial —determinante-pronombre— se corrijan configuraciones válidas como *La mostaza la echo al final.*

Por otro lado, la versión plural **No acepto tu análisis de los echos,* quedaría corregida por el patrón general que se presentó al comienzo de este epígrafe.

11.2.8 Errores que involucran la locución *de hecho*

Este error queda cubierto y corregido por el siguiente patrón de error:

`de echo`	`[- echo]` `[+ hecho]`	La locución *de hecho,* que suele acompañarse de comas, se escribe con h- inicial y tiene como significado *efectivamente, de veras, con eficacia y buena voluntad*. Además, en derecho, significa *sin ajustarse a una norma.* *No pienso volver a escribirte, de hecho he borrado todos tus datos de contacto.*

11.3 *deshecho* y *desecho*

Otro de los errores recurrentes que emergen del fenómeno de la homofonía surge por la confusión entre el sustantivo masculino *desecho,* las variantes participiales del verbo *deshacer* y algunas formas del presente de indicativo del verbo *desechar*; *desecho, desechas, desecha.*

Para atajar esta vacilación entre homófonos ha sido necesario establecer un subgrupo de 74 patrones de error que abarquen todas las posibilidades formales y entornos sintácticos en los que estas formas pueden participar.

Como en el caso de los pares *hecho-echo* y sus derivados, cada patrón codificado solventará, de modo parcial, errores provocados por la concurrencia de alguna de estas formas. Conviene, para llevar a cabo esta tarea gradual, tener acotados los contextos y la distribución propia de cada una de estas palabras.

11.3.1 Contextos de las formas participiales de *deshacer*

Las formas participiales del verbo *deshacer* están presentes en dos configuraciones de naturaleza verbal. En primer lugar, la forma masculina y singular, *deshecho,* es requerida para la formación de los tiempos verbales compuestos del verbo; *habíamos deshecho, hubiera deshecho, habré deshecho*, etc. Junto a esta, las formas restantes *deshecha, deshechas* y *deshechos* son necesarias para la construcción de perífrasis de participio en las que participa este verbo; *estar deshecho, llegar deshecha, quedar deshechas,* etc.

Este participio, por ser propio de un verbo transitivo[91], puede funcionar como adjetivo que complemente a un sustantivo. Ejemplos de estos usos adjetivos son los sintagmas *trenzas deshechas, trabajo deshecho, esperanzas deshechas,* etc.

11.3.2 Contextos de las formas flexionadas de *desechar*

Las formas conjugadas de la primera, segunda y tercera persona del singular del presente de indicativo del verbo *desechar* se solapan con las variantes participiales de *deshacer* al ocupar los mismos dominios oracionales. En la lengua escrita, como cabe esperar, su ortografía es causa de múltiples errores. Es un error frecuente, quizá por asimilación, la inclusión en la grafía del resto de las formas de la conjugación de *desechar* la *h-* propia del verbo *deshacer*.

[91] Este fenómeno de *desdoblamiento* de un participio en adjetivo deverbal ha resultado ser inviable con algunos verbos intransitivos de naturaleza inergativa. Sirvan como ejemplo de esta imposibilidad: *reír*, ?*El reído chiste; vagar*, ?*Las vagadas calles; oscilar*, ?*Su oscilada respuesta.*

11.3.3 Contextos del sustantivo *desecho*

La forma *desecho,* además, se presta a ser un sustantivo masculino con el sentido de *residuo despreciable por no tener utilidad*.

Del mismo modo que otros sustantivos podrá ser el núcleo de un grupo nominal, por lo que generalmente irá precedido de un determinante que lo habilite como núcleo. Otra de sus configuraciones frecuentes es el sintagma preposicional *de desecho* en el que el sustantivo funciona como término.

Como en el epígrafe anterior, se presentarán, organizadas en baterías de patrones, las medidas correctivas que se han diseñado a partir de los datos que nos aportan el corpus sobre los contextos típicos de estas formas. Cada patrón solventará errores parciales y, yuxtapuesto a otros patrones, ofrecerán cobertura a la totalidad compleja de este fenómeno.

11.3.4 Errores que involucran formas participiales de *deshacer*

Para solventar los errores derivados de la ortografía del participio **desecho* en las formas compuestas se ha codificado un único patrón que aborda toda la casuística.

`#haber desecho`	`[-desecho][+deshecho]`

El resto de las estructuras en las que pueden participar las cuatro formas participiales del verbo *deshacer* se encuentran en las perífrasis de participio; *estar deshecho, quedarse deshechas, volver deshechos*, etc.

Aunque, intuitivamente, parecería viable la codificación de un patrón genérico como `#{vb} desecho` que corrigiera la forma de participio, una vez más la ambigüedad obliga a condicionar esta regla. Si se expandiera el patrón a cualquier verbo, expresiones que se construyen con el sustantivo *desecho* como *para desecho* o *entre desechos*, quedarían corregidas —*desecho* por *deshecho*— al tratar las preposiciones *entre* y *para* como formas conjugadas de *entrar* o *parar*. Por otro

lado, es correcta la combinación del sustantivo *desecho* con algunos verbos como *considerar, llamar, recoger* o *reciclar desechos.* Estas secuencias, ante la aplicación de una regla genérica como la sugerida, quedarían corregidas erróneamente.

Ante estas perspectivas, se ha optado por codificar patrones individuales con los verbos que se ha comprobado en el Corpus TIP que con mayor frecuencia coaparecen en perífrasis cuyo participio sea una variante de *deshacer.* El modelo que se ha seguido para la formalización de estos errores es el que sigue:

`#estar @{adv}+? desecho`	**Estaba tan desecho como pensaba.*
`#estar @{adv}+? desechos`	**Han estado muy desechos durante la ceremonia.*
`#estar @{adv}+? desecha`	**Está desecha de dolor.*
`#estar @{adv}+? desechas`	**Las galletas estaban desechas por los golpes.*

Los verbos que por su frecuencia se han codificado en patrones de perífrasis son los siguientes:

`ser`	`volver`	`llegar`
`quedar`	`venir`	`aparecer`

Otro de los contornos y funciones que pueden asumir las formas de participio se observa en los sintagmas nominales en los que el núcleo está complementado por el adjetivo deverbal `@deshecho`. Para dar cobertura a las expresiones que presenten esta configuración propia de los adjetivos se codifica el siguiente bloque de patrones en los que se aplicará la misma acción correctiva; la sustitución de la secuencia *desech-* por *deshech.*

`@{N}@@{masc sg} desecho`	**hielo desecho*
`@{N}@@{fem sg} desecha`	**mujer desecha*
`@{N}@@{masc pl} desechos`	**puzles desechos*
`@{N}@@{fem pl} desechas`	**ilusiones desechas*

Estos patrones están sometidos a la condición de que el sustantivo no sea *de, pero* y *por.* Estas formas, que son ambiguas —preposición-conjunción y sustantivo— corregirían secuencias correctas como *Se compran materiales de desecho; Te gusta la persona, pero desechas sus propuestas,* etc.

Por último, y dentro del desempeño de estas funciones adjetivales, se cierra el tratamiento de los errores derivados de la escritura del participio de *deshacer* con un grupo de patrones encabezados por un adverbio modificador; *muy deshecho, absolutamente deshechas, bien deshechos*, etc.

Una vez más, y ante la tentación de llevar a cabo generalizaciones, la aplicación de la regla debe ceñirse a unos cuantos adverbios para evitar falsos positivos como; *No desechas los malos hábitos; Prácticamente desecho todas sus propuestas, Nunca desechas nada de tu armario.*

El modelo de codificación que sigue se ha aplicado a otros adverbios que participan de algunas de las expresiones más frecuentes.

bien desecho	**El compuesto debe echarse bien desecho.*
bien desecha	**La papilla le gusta bien desecha.*
bien desechos	**Llegaron de la excursión bien desechos.*
bien desechas	** Las galletas bien desechas se colocarán como base de la tarta.*

mal	tan	absolutamente
muy	totalmente	prácticamente

11.3.5 Errores que involucran formas flexionadas de *desechar*

Como en el caso del verbo *echar,* las interferencias con los participios de *hacer*, y en este caso de *deshacer,* generan errores y vacilaciones con las formas del verbo *desechar.* Es habitual, por lo tanto, encontrar este verbo con *h-*, y no solo en las formas homófonas que se solapan con las de *deshacer* sino en el resto de formas;

**deshecharías, *deshecharemos, *deshecharán,* etc. Se llevan a cabo, para solventar estos errores, dos líneas de actuación.

Por un lado, se han codificado dos reglas generales restringidas que encapsulan el grueso de la casuística que deriva del uso del verbo **deshechar* y su variante pronominal **deshecharse*. La primera medida que ha de tomarse es la creación de la conjugación de los verbos, ambos inexistentes en el lexicón. De este repertorio generado habrá que extraer las formas que coincidan con los participios *deshecho, deshecha, deshechos, deshechas,* para que el programa no corrija automáticamente secuencias correctas desarrolladas por el participio del verbo *hacer.*

`#deshechar`	`[-deshe][+dese]`

Junto a estas instrucciones generales, se codificarán patrones más acotados que ayuden a cercar este tipo de errores y ofrezcan tratamiento para las formas homófonas que han quedado excluidas. Para facilitar la identificación del error por parte del programa ha sido necesario, en el diseño del patrón, apoyarse en dos posibles cotextos del margen izquierdo:

`@{pn}@@{sujeto} no? #deshechar`	**Tú deshechas mis consejos, así te va.*
`@{pn}@@{reflex} #deshechar`	**La propuesta se deshecha por sí sola.*

En el caso del último patrón en el que intervienen los pronombres reflexivos será necesario extraer aquellos pronombres masculinos que resulten ambiguos con los artículos. Se evitan, así, falsos positivos como *Los* (polvorones) *deshechos no hay quien se los coma*.

11.3.6 Errores que involucran al sustantivo *desecho*

Se aborda, por último, el tratamiento del sustantivo masculino *desecho* mediante el siguiente bloque de patrones que recogen los usos y expresiones que han resultado ser más frecuentes según los resultados arrojados por el corpus:

`@{det}@@{sg}@@{masc} deshecho`	**El deshecho nuclear preocupa a la población.*
`@{det}@@{pl}@@{masc} deshechos`	**Los deshechos de la fábrica siguen perjudicando al ecosistema de la zona.*
`@deshecho @tóxico`	**El asunto del deshecho tóxico abrió todas las noticias.*
`de deshecho`	**Los materiales de deshecho requieren un tratamiento específico.*

En todos los casos se aplica el mismo tratamiento sobre el error; sustitución del segmento *deshech-* por *desech-.*

11.4*ahí*, *ay*, *hay*

Una tripla clásica dentro del universo de los homófonos es la protagonizada por *hay, ahí, ay.* Podemos acotar estas formas vacilantes con recursos válidos para la identificación como su categoría gramatical y su distribución. Básicamente, la forma *hay* corresponde a una forma terciopersonal del verbo *haber*. Sus dominios y configuraciones serán los propios de un verbo impersonal en tercera persona del singular. *Ahí*, por su parte, es un adverbio de lugar que puede ocupar múltiples entornos y, por último, la forma *ay,* que suele aparecer encorsetada entre exclamaciones, es una interjección utilizada para expresar *muchos y muy diversos movimientos del ánimo* (DRAE).

Para el tratamiento de los errores que estas formas provocan en la escritura se han codificado 17 patrones, 14 de ellos detectarán errores de escritura que involucran estas formas y las corregirán automáticamente. Los tres patrones restantes, por su parte, están diseñados para actuar como avisos lingüísticos que aporten información al usuario sobre la naturaleza y uso de estas formas. Se

activarán solo en el caso de que estas formas no sean reconocidas previamente por los patrones etiquetados como error.

A continuación, se expone el trabajo que se ha llevado a cabo para abordar este asunto.

11.4.1 Errores que involucran el adverbio *ahí*

La siguiente batería recoge patrones cuyas configuraciones y formas constituyen inequívocamente errores por presentar las formas *hay* o *ay* en lugar del adverbio. La codificación se ha desarrollado a partir del estudio de los cotextos posibles que acompañan a este adverbio y son, por su distribución, privativos de otras las otras formas.

`@{pn reflex}? #{vb} hay`	`[- hay][+ ahí]`	**Se compran hay.*
`hay @{pn reflex}? #{vb}`	`[-hay ][+ahí ]`	**Hay se vive fenomenal.*
`@{pn reflex}? #{vb} ay`	`[- ay][+ ahí]`	**Nos vamos ay.*
`ay @{pn reflex}? #{vb}`	`[-ay ][+ahí ]`	**Ay se casaron.*
`de hay`	`de ahí`	**Es hijo único, de hay su conducta.*
`de ay`	`de ahí`	**Es muy infantil, de ay que no haya que hacerle caso.*
`hay #{vb}`	`[-hay ][+ahí ]`	**Hay van aquellas mujeres alocadas.*
`ay #{vb}`	`[-ay ][+ahí ]`	**Ay quería yo llegar.*

11.4.2 Errores que involucran el verbo *hay*

En este bloque se abordan los casos en los que una secuencia constituye un error por presentar, con ese cotexto, otras variantes homófonas distintas de *hay.*

`@{pn de CD} ahí que`	`[-ahí que][+hay que]`	**Los ahí que llaman desesperados.*
`ahí que {inf}`	`[-ahí que][+hay que]`	**Ahí que retrasar el envío.*
`@{pn de CD} ay que`	`[-ay que][+hay que]`	**Las ay que beben al volante.*
`ay que {inf}`	`[-ay que][+hay que]`	**Ay que dejarlo todo recogido.*
`@{adv} ahí`	`[- ahí][+ hay]`	**No ahí otra alternativa.*
`@{adv} ay`	`[- ay][+ hay]`	**Aquí ay más soluciones.*

11.4.3 Aviso lingüístico

Como se señaló, el subproceso que lleva a cabo el programa consiste en la búsqueda, en el texto objeto de corrección, de secuencias que se identifiquen con los patrones de error codificados y, en caso de no coincidir con ninguno de ellos, aplicarle el patrón de aviso, que se activará ante la identificación de cualquiera de las formas *ahí, ay* o *hay.*

> Para hacer referencia a la forma del verbo *haber* que indica la existencia de algo, la forma correcta es *hay*: *Hay mucho por hacer*.
>
> Para aludir al adverbio de lugar (sustituible por *allí*), la forma correcta es *ahí*: *Ahí/allí compré aquellos tomates tan ricos*.
>
> La forma *ay* es una interjección utilizada para expresar diversas emociones (dolor, sorpresa, susto) y suele ir acotada entre exclamaciones y/o seguida de la preposición *de*: *¡Ay de mí!*, *¡Ay del que se atreva a tocarte!*

11.5 Casi homófonos

Se abordarán, en este epígrafe, aquellos pares o grupos de palabras que por su naturaleza casi homófona suscitan dudas y con frecuencia inducen a errores en el proceso de su escritura.

Es común, en la lengua oral, que la pronunciación de palabras que contienen los sonidos /ll/, /x/, /cc/ se relaje y actualice en sonidos más laxos; /y/, /s/, /c/. Esta relajación e indiferenciación de los sonidos, que tiene repercusión en la escritura genera, de facto, una neutralización de la oposición que mantenían estos términos, dando lugar a fenómenos de homofonía en términos que en principio no la padecían.

Como puede intuirse, este ámbito de los homófonos y casi homófonos participa de cuestiones tanto ortográficas como léxico-semánticas. No obstante, solo captando estas últimas nociones y la intención significativa con que se utiliza una palabra puede ofrecerse una ortografía correcta. Así, ante la dicotomía *arroyo*-

arrollo, solo el conocimiento de información léxica y categorial, que inmediatamente educe un entorno sintáctico, puede resolver la disyuntiva ortográfica, habida cuenta de que todas las formas tratadas son posibles y están presentes en el lexicón.

Se exponen a continuación, microrepertorios de formas que han sido objeto de tratamiento para la corrección de errores frecuentes derivados de términos casi homófonos.

Dividiremos este contenido en dos bloques; en primer lugar, se ofrece una muestra de cómo se han tratado y codificado algunos errores que provocan estos términos en la escritura. En el segundo bloque se expondrán los avisos lingüísticos que aportan información de carácter léxico y se le ofrecen al usuario cuando el programa detecta alguna de estas formas.

11.5.1 Errores

Pares como *raya-ralla, arroyo-arrollo* o *adición-adicción* junto con sus derivados, son algunos de los casos de casi homófonos que con frecuencia dan lugar a errores ortográficos en la escritura. De la masa de datos que ofrece el corpus de frases se han extraído sus cotextos certeros y se han codificado en patrones los errores más plausibles que estas formas pueden protagonizar.

`@{N} @rayado`	`[- rayad][+ rallad]`	**huevo rayado*
`con rallas`	`con rayas`	**camiseta con rallas*
`a rallas`	`a rayas`	**pijama a rallas*

`#sufrir @{det}+? @adición`	`[-adici]` `[+adicci]`	**Sufría adiciones muy severas.*
`#padecer @{det}+? @adición`	`[-adici]` `[+adicci]`	**Padecía la adición a la heroína de su hijo.*

#arroyar	[-arroy] [+arroll]	*Arroyaron a un ciclista esta mañana.
@{pn reflex}? arroyo	[-arroyo] [+arrollo]	*Te arroyo si no me dejas pasar.
@{det}@@{masc}+ @{adj}? arrollo	[-arrollo] [+arroyo]	*Los claros arrollos quedaron perpetuados en mi memoria.

Debe advertirse que para la consecución de las acciones del primer patrón es necesaria la generación previa de la conjugación del verbo ficticio **arroyar*, que por ser una forma errónea no está presente en el Lexicón TIP.

Para su aplicación efectiva tiene como condición la exclusión de la primera persona de singular del presente de indicativo, *arroyo*, con el fin de evitar correcciones indebidas sobre el sustantivo.

Por último, se ofrece un patrón para el que, como en el caso anterior, es necesaria la generación previa de la conjugación del verbo erróneo **cayar*. La condición que se impone en este caso es la exclusión de este repertorio de las formas ambiguas como *cayo, cayó, cayado*, etc. para evitar correcciones indebidas sobre estructuras correctas.

#cayar	[-cay][+call]	*No os cayásteis en la reunión de vecinos.

11.5.2 Avisos lingüísticos

Los siguientes avisos se ofrecerán ante la identificación en el texto de pares de palabras que, por la mencionada pronunciación relajada de algunos sonidos del español, generan dudas e interferencias tanto el plano semántico, como en el ortográfico; *esotérico-exotérico; rayar-rallar; espirar-expirar.*

Junto con el patrón de error que suscita el aviso lingüístico, se aporta el contenido informativo que el programa ofrece al usuario:

#tener @{det}+? @adición[92]	Para referirse al *hábito de quien se deja dominar por el uso de alguna o algunas drogas tóxicas, o por la afición desmedida a ciertos juegos* la forma correcta es *adicción*. *Adición*, por su parte, se refiere a la *operación de sumar, añadidura que se hace*. Aunque ambas palabras pueden sonar igual –son casi homófonas–, su grafía y significado son diferentes.

@arroyo	Este sustantivo se refiere a un *caudal corto de agua*. Para hacer referencia a la 1ª pers. de sg. del verbo *arrollar* con el significado de *atropellar*, la forma correcta es *arrollo*.
arrollo	Esta forma es la 1ª pers. de sg. del verbo *arrollar*, que significa *atropellar*. Para hacer referencia al *caudal corto de agua*, la forma correcta es *arroyo*.

calló	Esta forma es la 3ª pers. de sg. del pretérito perfecto simple del verbo *callar*. Para hacer referencia al verbo *caer*, la forma correcta es *cayó*.
cayó	Esta forma es la 3ª pers. de sg. del pretérito perfecto simple del verbo *caer*. Para hacer referencia al verbo *callar*, la forma correcta es *calló*.

#escavar	Este verbo significa *cavar ligeramente la tierra para ahuecarla*. Con una pronunciación relajada esta forma tiene un homófono, *excavar*, cuyo significado es *hacer en el terreno hoyos, zanjas, desmontes, pozos o galerías subterráneas*. Aunque ambas palabras pueden sonar igual —son casi homófonas—, su grafía y significado son diferentes.
#excavar	Este verbo significa *hacer en el terreno hoyos, zanjas, desmontes, pozos o galerías subterráneas*. Con una pronunciación relajada esta forma tiene un homófono, *escavar*, cuyo significado es *cavar ligeramente la tierra para ahuecarla*. Aunque ambas palabras pueden sonar igual —son casi homófonas—, su grafía y significado son diferentes.

@esotérico	Este adjetivo significa *oculto, reservado, impenetrable o de difícil acceso para la mente*. Con una pronunciación relajada esta forma tiene un homófono, el adjetivo *exotérico*, cuyo significado es contrario: *común, accesible para el vulgo, de fácil acceso para la mente*. Aunque ambas palabras pueden sonar igual, son casi homófonas, su grafía y significado es diferente.
@exotérico	Este adjetivo significa *común, accesible para el vulgo, de fácil acceso para la mente*. Con una pronunciación relajada esta forma tiene un homófono, el adjetivo *esotérico*, con un significado contrario: *oculto, reservado, impenetrable o de difícil acceso para la mente*. Aunque ambas palabras pueden sonar igual, son casi homófonas, su grafía y significado es diferente.

[92] Esta forma se codifica con el cotexto izquierdo con el fin de acotar el aviso a una situación susceptible de contener un error. De no acotar la forma, el aviso se activaría en cualquier circunstancia ante la secuencia, pudiendo llegar a ser molesto para el usuario.

#espirar	Este verbo significa *expulsar el aire de los pulmones*. Con una pronunciación relajada esta forma tiene un homófono, el verbo *expirar*, cuyo significado es *acabar la vida o un periodo de tiempo*. Aunque ambas palabras pueden sonar igual, son casi homófonas, su grafía y significado es diferente.
#expirar	Este verbo significa *acabar la vida o un periodo de tiempo*. Con una pronunciación relajada esta forma tiene un homófono, el verbo *espirar*, cuyo significado es *expulsar el aire de los pulmones*. Aunque ambas palabras pueden sonar igual, son casi homófonas, su grafía y significado es diferente.

@estatico	Este adjetivo significa *que permanece en un mismo estado, sin mudanza en él.* Con una pronunciación relajada esta forma tiene un homófono, el adjetivo *extático*, cuyo significado es *que está en éxtasis o lo tiene con frecuencia o habitualmente*.
@extatico	Este adjetivo significa *que está en éxtasis o lo tiene con frecuencia o habitualmente*. Con una pronunciación relajada esta forma tiene un homófono, el adjetivo *estático*, cuyo significado es *que permanece en un mismo estado, sin mudanza en él*.

#rallar	Este verbo significa *desmenuzado con un rallador* o, en un registro coloquial, *molestar, fastidiar con importunidad y pesadez*. Para hacer referencia a *hacer o tirar rayas*, el verbo correcto es *rayar*.
#rayar	Este verbo significa *hacer o tirar rayas*. Para hacer referencia a *desmenuzado con un rallador* o, en un registro coloquial, *molestar, fastidiar con importunidad y pesadez*, el verbo indicado es *rallar*.

11.6 Ortografía de algunas palabras

Se han codificado específicamente algunas de las palabras que con mayor frecuencia presentan error. Su especificación en un patrón pretende ahorrar al sistema la búsqueda de palabras similares y el cálculo de probabilidades necesario y previo a la ejecución de la corrección ortográfica automática. La identificación de este patrón supone la sustitución directa de la forma afectada lo que aligera el procedimiento de corrección que debe asumir el programa.

discursiones	discusiones
discursión	discusión
disgresiones	digresiones
disgresión	digresión

cortacircuito	cortocircuito
#desvastar	[-desvast][+devast]
#diverger	[-#diverger][+#divergir]
#transplantar	[-transplant][+trasplant]

Nótese que para los últimos tres patrones, cuya forma involucrada es un verbo, tendrán que ser generados malformados automáticamente.

Con el mismo afán de facilitar al autómata el tratamiento de estos errores, se han registrado algunas formas verbales que presentan una morfología irregular por la alternancia vocálica que entre *e*; *ié.* Estas formas que suelen ser portadoras de error se han incorporado manualmente como patrones de error individuales.

apreta	aprieta
apretan	aprietan
apretas	aprietas
aprete	apriete
apreten	aprieten
apretes	aprietes
apreto	aprieto

Esta batería se ha desarrollado de igual modo para otros verbos como:

restregar estregar plegar mentar fregar

A diferencia de los otros casos presentados que suelen pasar desapercibidos para los correctores, estas formas si son identificadas como error por otros programas como *Word* y reciben una corrección automática adecuada.

11.7 Palabras que admiten dos grafías

Se aborda, por último, el tratamiento de algunas formas que admiten dos posibilidades para su escritura; *ológrafo/hológrafo; psiquiatra/siquiatra*. La

corrección de ambas no obsta para que la norma, por diversos motivos según el caso, recomiende una grafía en lugar de la otra.

El siguiente repertorio de patrones contiene la versión desaconsejada de algunos de los pares más frecuentes que han sido extraídos de las obras de referencia que sirven como fuente de datos de PatErr. La sustitución de las formas del patrón por las recomendadas es una ejecución que el usuario podrá automatizar.

@magazín	[-magaz][+magac]
@zirconita	[-zirc][+circ]
@alelí	[-ale][+alhe]
@ambidextro	[-ambidext][+ambidiest]
@asera	[-aser][+acer]
@azimut	[-azim][+acim]
@bacón	[-bac][+beic]
@baraúnda	[-baraúnd][+barahúnd]
@bechamel	[-bech][+bes]
@bolladura	[-bollad][+abollad]
@calina	[calin][+calim]
@cantilena	[-cantilen][+cantinel]
@carrillón	[-carrill][+carill]
@cayak	[-cayak][+kayac]
@cazcarria	[-cazcarr][+cascarr]
@chavola	[-chav][+chab]
@genízaro	[-genízar][+jenízar]
@gineta	[-ginet][+jinet]
@hacera	[-hace][+ace]
@harem	[-harem][+harén]
@harmonía	[-harmon][+armon]
@harpa	[-harp][+arp]
@higuana	[-higuan][+iguan]
@hológrafo	[-hológraf][+ológraf]
@hurraca	[-hurra][+urra]
@jiga	[-jig][+gig]
@kamikace	[-kamikac][+kamikaz]
@mnemotecnia	[-mnemotecn][+nemotecn]
@nomo	[-nom][+gnom]

@nóstico	[-nósti][+gnósti]
@piyama	[-piyam][+pijam]
@quirie	[-quir][+kir]
#rembolsar	[-#rembolsar][+#reemplazar]
#remplazar	[-rempl][+reempl]
@sicología	[-sicolog][+psicolog]
@sicólogo	[-sicólog][+psicólog]
@sicópata	[-sicóp][+psicóp]
@sifonier	[-sifon][+chifon]
@siquiatra	[-siquiat][+psiquiat]
@vargueño	[-varg][+barg]
@vodca	[-vodc][+vodk]
@wolframio	[-wolfram][+volfram]
@yedra	[-yed][+hied]
@yerba	[-yer][+hier]
@zebra	[-zebr][+cebr]
@zedilla	[-zedi][+cedi]
@zenit	[-zen][+cen]
@zigoto	[-zigo][+cigo]
carst	karst
@ázimo	[-ázim][+ácim]
@benzina	[-benz][+benc]
@ceta	[-cet][+zet]
@zinc	[-zin][+cin]
@zíngaro	[-zíng][+cíng]

Capítulo 12

Morfología

Las soluciones a los problemas relacionados con la morfología que aquí se abordan, se basan, en buena parte, en el sistema de listas que es capaz de ofrecer un tratamiento unificado a palabras que por algunos rasgos formales conviven en un mismo paradigma —de los varios posibles en los que pueden participar las palabras—.

Se recogerán, a continuación, aquellos aspectos relativos a la flexión de número de ls sustantivos que suelen pasar desapercibidos para el escritor y dan lugar a error. Debe reconocerse que el tratamiento de estos temas es abordado por otros sistemas de verificación con un amplio grado de acierto y cobertura. No obstante, el repertorio de errores de PatErr se asocia, como es sabido, a glosas lingüísticas que se le ofrecen al escritor para que tome conciencia de las causas morfológicas o excepciones que han provocado ese error.

Con el fin de lograr una gestión unificada de los contenidos, algunos de los errores derivados de procesos morfológicos se abordan, contextualizados, en otros capítulos relacionados con estos aspectos, como el dedicado a la ortografía o, dentro de la gramática, a la concordancia.

12.1El plural de algunos sustantivos compuestos

El procedimiento morfológico de composición de nuevos términos es un campo en el que, con frecuencia, se observan reglas quebradas que dan lugar a error. Palabras como *salvamantel, espantapájaro, quitaesmaltes* o *guardabrisas*, constatan estos desvíos que se generan en el nivel morfológico.

La coordinación de lexemas como mecanismo de generación de una nueva forma nominal tiene ciertas restricciones que el escritor debe contemplar. Retomando los ejemplos expuestos en los que la composición deriva de la yuxtaposición de un verbo y un sustantivo, podemos vislumbrar dos grupos que se rigen por reglas opuestas.

Por un lado, sustantivos compuestos como **paragua*, **salvamantel* o **espantapájaro* exigen el plural como número inmanente; por otro, **quitaesmaltes* o **guardabrisas* deben mantenerse en singular en cualquier caso.

De estas restricciones se deduce que cuando el sustantivo que se adhiere al verbo para rematar el nuevo término es contable debe hacerlo en plural; *abrelatas.* Por el contrario, si el nombre que forma el compuesto es no contable el género que deberá asumir es el singular; *quitaesmalte.*

Para resolver estos errores de modo automático se han creado sendas listas (23 y 24) que recogen, por separado, todos los términos compuestos cuya estructura es «Vb+N». Los datos contenidos derivan de la compilación manual de estos sustantivos que se han extraído de la bibliografía especializada y de los ejemplos aportados por las obras normativas.

{Lista 23}	[+s]	Los nombres compuestos por «Vb+N» se construyen con el segundo elemento en plural si el nombre es contable: *friegaplatos*, *sacapuntas*, *rompeolas*, etc. En este caso el sustantivo es contable por lo que el compuesto debe estar en plural.
paracaída	[-paracaída] [+paracaídas]	

Complementariamente, se codifica el siguiente patrón;

{Lista 24}s	[-s]	Los nombres compuestos por «Vb+N» se construyen con el segundo elemento en singular si el nombre no es contable: *cortacésped*, *crecepelo*, *quitaesmalte*, etc. En este caso el segundo elemento se considera incontable, por lo que debe aparecer en singular.
guardarropas	[-guardarropas] [+guardarropa]	

12.2 *Pluralia tantum*

Se denomina *pluralia tantum* a los sustantivos que solo deben emplearse en plural, es decir, su número es inherentemente plural. Algunos ejemplos de este tipo de sustantivos son *nupcias*, *víveres*, *entendederas*, *cosquillas* y *redaños*, y algunos adjetivos nominalizados como *rompecorazones*, *buscavidas* o *zampabollos*. Tras una búsqueda y localización de estos sustantivos se han desarrollado dos listas que contienen algo más de 50 términos.

El patrón que codifica el error por el que uno de los elementos de este elenco es utilizado con forma singular será el siguiente:

{det}@@{sg}+ {adj}@@{sg}? carie	[-{det}@@{sg} {adj}@@{sg}? carie] [+{det}@@{pl} {adj}@@{pl}? caries]
{det}@@{sg}+ {adj}@@{sg}? {Lista 10A}	[-{det}@@{sg} {adj}@@{sg}? {Lista 10A}] [+{det}@@{pl} {adj}@@{pl}? {Lista 10A}s]

{det}@@{sg}+ {adj}@@{sg}? andurrial	[-{det}@@{sg} {adj}@@{sg}? andurrial] [+{det}@@{pl} {adj}@@{pl}? andurriales]
{det}@@{sg}+ {adj}@@{sg}? {Lista 10B}	[-{det}@@{sg} {adj}@@{sg}? {Lista 10B}] [+{det}@@{pl} {adj}@@{pl}? {Lista 10A}s]

Para corregir de modo automático todos los términos que presentan esta característica en cuanto al número. Es necesario para ello, hacer una lista diferenciada para aquellos términos terminados en vocal que construyen su forma correcta con *-s*; *gárgaras, añicos, fauces*, etc. y otra para los que terminando en consonante requieren la vocal para flexionar, *-es*; *maitines*, *enseres*, *víveres*, etc.

Por otro lado, se han codificado los patrones dando cobertura a los determinantes y adjetivos —en singular— que puedan preceder al término de plural inherente. De este modo quedarán corregidos y unificados en cuanto al número todos los integrantes del sintagma en el que el núcleo sea uno de estos sustantivos.

12.3*Singularia tantum*

Complementariamente, se tratan aquellos términos que forman parte del grupo de palabras *singularia tantum* y presentan una forma inherentemente singular; *norte, caos, sed, salud,* etc.

El tratamiento de estas formas es paralelo al que le precede; a partir de un sistema de listas se ajusta, mediante la supresión de las marcas de plural, la forma correcta de estos términos.

`{det}@@{pl}+ {adj}@@{pl}? oestes`	`[-{det}@@{pl} {adj}@@{pl}? oestes]` `[+{det}@@{sg} {adj}@@{sg}? oeste]`
`{det}@@{pl}+ {adj}@@{pl}? {Lista 11A}`	`[-{det}@@{pl} {adj}@@{pl}? {Lista 11A}]` `[+{det}@@{sg} {adj}@@{sg}? {Lista 11A}-s]`

`{det}@@{pl}+ {adj}@@{pl}? cenites`	`[-{det}@@{sg} {adj}@@{sg}? cenites]` `[+{det}@@{pl} {adj}@@{pl}? cenit]`
`{det}@@{pl}+ {adj}@@{pl}? {Lista 11B}`	`[-{det}@@{pl} {adj}@@{pl}? {Lista 11B}]` `[+{det}@@{sg} {adj}@@{sg}? {Lista 10A}-es]`

Para el desarrollo de estas listas se han excluido sustantivos ambiguos para evitar correcciones indebidas como en el caso de *sedes* —plural correcto de *sede*, o incorrecto de *sed*— o *saludes* —flexión del verbo *saludar*—. Aunque para este último caso podría parecer que el cotexto izquierdo podría servir de criba para seleccionar exclusivamente la variante nominal, secuencias verbales como *Las saludes*, podrían ser corregidas indebidamente tratando el pronombre como un determinante y el verbo como un sustantivo en plural.

12.4Sustantivos terminados en *-í*, *-ú*

Los plurales de sustantivos acabados en *-í* y *-ú* pueden construirse con *-es* o *-s*, aunque la norma culta recomienda la construcción de estos plurales con *-es*, optando por *bisturíes* en lugar de *bisturís* y *bambúes* en lugar de *bambús*.

Como casi todas las reglas, esta también presenta su excepción que afecta a algunas formas de las terminadas en *-u*; el plural de *champú, cucú, interviú, menú, tutú* y *vermú* se prefiere con la versión reducida en *-s*.

Para el tratamiento unificado de estos sustantivos se han desarrollado dos listas que recogen las formas terminadas en *-i* con un plural en *-s* —Lista 14—, y otra similar con los acabados en *-ús* —Lista 15—. De esta última se extraen las excepciones que han quedado expresas. Los patrones diseñados para el tratamiento de estos plurales son los siguientes:

@{Lista 14} alhelís	[-ís][+íes]	Los nombres acabados en *-í* y *-ú* se construyen con *-es* o *-s*, pero la norma culta recomienda la primera opción: *bisturíes*. La excepción a esta regla la constituyen las siguientes formas terminadas en *-ú*: *champús, cucús, interviús, menús, tutús* y *vermús*.
@{Lista 15} iglús	[-ús][+úes]	

12.5 Sustantivos terminados en *-y*

Precedida de vocal

Sustantivos como *buey, virrey* o *convoy* cuya terminación tiene forma de vocal seguida de *-y* construyen su plural añadiendo *-es*. Además de los propios de nuestra lengua, se rigen por esta norma otros nombres foráneos plenamente castellanizados como *convoy* o *carey*. No sucede así con algunos sustantivos con esta misma configuración como *jersey, gay, espray* o *yóquey*, en los que, por su reciente incorporación en nuestro repertorio, la *-y* del singular mantiene en plural su carácter vocálico y en consecuencia se transforma en *-i* para formar el plural. Así se prefieren las formas de plural *gais*, *jerséis, espráis* y *yoquis*.

Como en los casos precedentes, se ha recopilado una lista —35— con todos los términos en singular que cumplen con estas características y se le adherido una *-s* para formar su plural. Se han excluido los mencionados términos que contradicen esta norma.

El tratamiento de estos errores derivados de la flexión de número se basa en el siguiente patrón:

@{Lista 35} *ays*	[-ys][+yes]	El plural de esta forma terminada en *-y* precedida de vocal, así como de aquellas extranjeras plenamente castellanizadas, *convoy* o *carey*, se construye añadiendo *-es*: *Se escucharon los ayes desde el otro lado del río*.

Precedida de consonante

Algunos sustantivos extranjeros terminados en *-y* precedida de consonante se adaptan gráficamente al español sustituyendo la *-y* por *-i*. Este cambio, lógicamente, estará presente en su plural: *dandis* o *pantis*. A pesar de esta regla de adaptación a nuestra lengua, son muchas las ocasiones en las que estas palabras se escriben con *-y* en singular; *hobby*, *derby*, etc.

Para subsanar estos errores se ha compilado una lista de palabras terminadas en *-y* precedida de una consonante a las que se le ha añadido una *-s* para formar su plural. El patrón diseñado es como el que sigue:

@{Lista 22} *ferrys*	[-ys][+is]	El plural de esta forma extranjera terminada en una *-y* que está precedida de una consonante (*derby, dandy,* etc.) se construye añadiendo *-s*: *No he estado en ninguno de los derbis de la capital.* La norma recomienda siempre la grafía con *-i* independientemente del número que adopte la forma.

12.6Sustantivos sin flexión de plural

Algunas expresiones de tratamiento como *fray, sor, san* o *don* permanecen invariables en plural. Para ofrecer una solución a los errores que se puedan derivar de estas formas se codifican los siguientes patrones:

frayes	fray
frays	fray
sores	sor
sanes @{N}@@{propio}	[-sanes][+ san]
dones @{N}@@{propio}	[-dones][+ don]

Para los dos últimos patrones que contienen formas ambiguas ha sido necesario codificar el cotexto derecho para evitar correcciones sobre las versiones verbales de *donar* y *sanar* que también desarrollan estas formas.

Por otro lado, un grupo reducido de sustantivos de articulación compleja también permanecen indolentes a las marcas de plural. Todos ellos, junto con un *-s* adherida, se recogen en la Lista 23.

{Lista 23}	[-ts] [+t]	Esta palabra se mantiene invariable para su forma de plural. Les sucede lo mismo a las formas *compost, karst*, *kibutz, test*, *trust*, que no reciben marcas de plural para evitar una secuencia de difícil articulación en español.
kibutzs	kibutz	

Junto a los sustantivos de esta lista, se ha desarrollado algún aviso lingüístico —ante el riesgo que conlleva la ejecución de un patrón de error— para otros términos que entrañan dudas o errores en su uso en plural. Sirva como ejemplo el siguiente caso:

extras	**Las calidades extras de nuestros jamones son nuestra mejor credencial.*	Según las orientaciones de la norma académica *extra* es invariable cuando significa *superior.* Podrá flexionar en plural cuando signifique *adicional*; *horas extras*. Cuando hace referencia al sustantivo, sigue la regla general; *Los extras de Romeo y Julieta cobraban un dineral.*

12.7 Algunos monosílabos

Las vocales

Los monosílabos que dan nombre a las vocales se rigen por la regla general de composición del plural en monosílabos, esto es, añadiendo *-es*. Se aparta de esta regla el sustantivo que se refiere a la letra *e* que forma el plural como los nombres de las consonantes en *-s.*

Para ofrecer tratamiento correcto a estos monosílabos se han diseñado los siguientes patrones:

`@{det}@@{pl}@@{fem}+ @{adj}? as`	`[- as][+ aes]`
`@{det}@@{pl}@@{fem} @{adj}? is`	`[- is][+ íes]`
`@{det}@@{pl}@@{fem} @{adj}? ís`	`[- ís][+ íes]`
`@{det}@@{pl}@@{fem} @{adj}? os`	`[- os][+ oes]`
`@{det}@@{pl}@@{fem} @{adj}? us`	`[- us][+ úes]`
`@{det}@@{pl}@@{fem} @{adj}? ús`	`[- ús][+ úes]`

Para facilitar la identificación de estos términos y evitar la intervención sobre secuencias homófonas como el sustantivo singular *as*, o el pronombre de tercera persona *os*, se han codificado los patrones contando con su cotexto izquierdo —femenino y plural— y abriendo la posibilidad de que entre el determinante y el sustantivo se intercale algún adjetivo; *Las inmensas aes del letrero luminoso estaban fundidas.*

Partículas de polaridad

`sís`	`síes`	La norma recomienda la construcción del plural de este monosílabo en *-es*: *síes*.
`@{det}@@{pl}@@{masc} @{adj}? nos`	`[- nos]` `[+ noes]`	La norma recomienda la construcción del plural de los monosílabos terminados en *-o*, en *-es*: *noes*. La excepción a esta regla es la forma *pro/pros*.

A propósito de este último patrón, se diseña otro para abordar el tratamiento correcto de la construcción del plural en los monosílabos terminados en *-o*, que debe hacerse mediante la adhesión del morfema *-es*; *yoes*. La excepción a esta regla es la forma *pro*, cuyo plural será *pros*.

12.8Algunas expresiones compuestas

Expresiones de uso reciente como el plural de *corta y pega* generan dudas en cuanto a qué elemento debe desarrollar la flexión de número. Esta secuencia formada por dos verbos no se ha lexicalizado aún, por lo que se recomienda no añadir marca de pluralidad aun cuando se refiera a varios *corta y pega*. No tienen esta restricción otras expresiones lexicalizadas que comparten configuraciones compuestas como *vaivenes, sabelotodos, quitaipones* o *correveidiles.*

El tratamiento de este tipo de secuencias complejas requiere la formalización de tres patrones de error que parcialmente sean capaces de copar toda la casuística que el error puede provocar.

corta y pegas	corta y pega
cortas y pega	corta y pega
cortas y pegas	corta y pega

Capítulo 13

Gramática

El abordaje de temas gramaticales a partir de técnicas de bajo nivel supone un reto arduo para el ingeniero lingüístico que se proponga este objetivo. Estas técnicas, como se ha dejado de manifiesto, analizan la lengua en un nivel superficial pero concreto, fácilmente formalizable para la máquina, a diferencia de las técnicas abstractas que derivan en análisis profundos de naturaleza sintáctica o semántica.

Pero solucionar ciertos casos en los que se viola una regla gramatical, aunque sea producto de un error de actuación —lapsus, error tipográfico, etc.—, precisa de un complejo sistema de reglas plagadas de matices que excede los límites de la propuesta que se actualiza en PatErr.

En cualquier caso, tras un estudio de los fenómenos gramaticales que con mayor frecuencia constituyen errores del español escrito, se han identificado parcelas inscritas en contextos locales en las que se desarrollan errores gramaticales que pueden ser tratadas a partir de patrones de error.

Por otro lado, varias de las incidencias aquí tratadas constituyen casos excepcionales o específicos para los que otros programas de revisión textual —*Word*, *Stilus* o *GramCheck*— no provén, por el momento, una solución.

Se ofrecerá, en este capítulo, el estudio y tratamiento que se ha desarrollado para algunas cuestiones que pertenecen al ámbito gramatical y suelen ser causa de errores en el texto escrito; asuntos relacionados con la concordancia, en todas sus actualizaciones, aspectos relacionados con la gramática verbal y, por último, otros fenómenos asociados a la rección preposicional, como la vacilación entre ellas, el dequeísmo o el queísmo.

13.1 Concordancia

Como sucederá con los asuntos referidos a la gramática verbal, la concordancia, entendida como sistema de relaciones que se establece entre las palabras en el eje sintagmático, requiere un análisis profundo de los constituyentes que va más allá del contexto inmediato de la palabra. La tarea de investigación que aquí se ha llevado a cabo, ha permitido extraer conclusiones, captar excepciones necesarias para matizar reglas, y desarrollar patrones "universales" que servirán en el futuro como base de un *motor de concordancia*[93].

El repertorio de temas investigados con respecto a estos asuntos se ha concretado en 321 patrones codificados. El contenido de estos ha sido granulado en ocho errores, dos recomendaciones y dos avisos diferentes, y puede ser organizado en el siguiente esquema simplificado:

I. Concordancia gramatical intrasintagmática: categorías gramaticales
II. Concordancia gramatical intersintagmático: relaciones entre sujeto y verbo
III. Concordancia de género
IV. Concordancia de número
V. Concordancia de persona

13.1.1 Concordancia intrasintagmática; categorías

En esta sección se abordan casos, generalmente ligados a expresiones, en los que se constatan problemas de concordancia gramatical generados por las propias restricciones que presentan las categorías gramaticales. La invariabilidad de los adverbios, el imperativo de concordancia de los determinantes, la ambigüedad categorial de algunas formas, etc., son aspectos que deben observarse y formalizarse para el desarrollo de soluciones relativas a la concordancia.

[93] Este, sin duda, es uno de los hitos más inmediatos que debe conquistarse con el fin de poder integrar este recurso en el marco de las industrias del lenguaje.

A continuación, se exponen, agrupados en dos bloques, los asuntos y fenómenos que han sido objeto de tratamiento en relación con discordancias en el entorno intrasintagmático.

13.1.1.1 *La forma* poco

Tras un estudio llevado a cabo sobre el corpus de frases, se ha observado que la forma ambigua *poco* es causante de varios errores de concordancia. Uno de los errores que se constatan con frecuencia queda representado en oraciones como **Compra una poca de carne; *Necesito una poca de agua.*

Este error, que pasa desapercibido para otros correctores, se genera cuando se intenta concordar el indefinido *poco* que, dentro de su ambigüedad categorial funciona aquí como un sustantivo —se puede asimilar a otros como *trozo, porción, kilo*, etc.— con el sustantivo núcleo del sintagma preposicional que le complementa.

Para solventar este error, se codifican dos patrones, en singular y en plural, del tipo:

```
una poca de @{N}@@{fem}
```

cuya acción correctiva se concreta en el cambio de género —de femenino a masculino— tanto del sustantivo, como de su determinante.

A propósito de este indefinido, se codifican otros casos en los que participa de expresiones erróneas por no aparecer seguido de la preceptiva preposición *de*: **Cómprame un poco queso.*

Se dedica otro patrón a este indefinido, en su versión adjetiva, que evita la interposición de la preposición *de* entre la secuencia *unos pocos* y el sustantivo al que cuantifica; **Aquel verano me leí unos pocos de libros*. Este uso arcaico es desaconsejable en el registro culto actual y como tal se codifica.

En relación con este último patrón, se registran otros tantos con el indefinido *cuanto*, que en ocasiones genera error cuando, en plural, aparece como en el caso anterior, con la preposición *de*. La norma exige, en estos contextos, eliminar la preposición ya que estas no son expresiones partitivas.

En síntesis, los patrones codificados para solventar estas incidencias motivadas por los indefinidos son los que siguen:

un poco @{N}	**Un poco leche*
una poca de @{N}@@{fem}	**Una poca de leche*
unos pocos de @{N}	**Unos pocos de días*
unas pocas de @{N}@@{fem}	**Unas pocas de veces*
unos cuantos de @{N}	**Unos cuantos de amigos*
unas cuantas de @{N}	**Unas cuantas de veces*

13.1.1.2 *Cuantificadores ambiguos*

Otra fuente de errores surgidos por motivos de concordancia es el uso de expresiones que contienen términos ambiguos categorialmente —adjetivo determinativo y adverbio, en este caso— del tipo *cuanto, mucho, bastante, igual*, etc.

Es frecuente encontrar en textos escritos oraciones como las siguientes;

**Cuanto más argumentos le doy, más reacio se muestra;*

**Los hoteles estaban muchos más llenos que el año pasado;*

**No eran los mejores preparados;*

**Aquella mujer estaba media mareada;*

**Aquellos vaqueros eran iguales de caros que estos.*

En estas oraciones se flexionan adverbios en busca de concordancias erróneas o se eluden concordancias forzosas entre determinantes y sus núcleos. El error, una vez más, surge por una vacilación entre las categorías de estas partículas ambiguas que según funcionen de una u otra manera establecerán relaciones —o no— con las secuencias que le siguen.

Los términos que con más frecuencia participan de este fenómeno —posibilidad de funcionar bien como adjetivo, bien como adverbio— y son causantes de errores de concordancia son los que siguen:

```
cuanto      poco        bastante    peor
mucho       medio       mejor       igual
```

Para todos ellos se han registrado patrones flexibles de error o aviso lingüístico, partiendo, en síntesis, del siguiente esquema básico:

término ambiguo (Adj - Adv)	**+ sustantivo**	**funciona como determinante**	**exige concordancia**
			Tuve bastantes motivos para actuar así.
	+adjetivo	**funciona como adverbio**	**rechaza concordancia**
			Los dos primeros tribunales fueron bastante más exigentes

En los patrones que incluyen estos términos se han codificado los dos posibles cotextos derechos; uno con sustantivo, en el que la corrección se basará en hacer concordar la partícula —que habrá de actuar como determinante— con su sustantivo, y el otro con adjetivo, donde el término ambiguo deberá mantenerse sin flexión como corresponde a los adverbios.

Se presenta un patrón modelo para ilustrar este planteamiento:

`cuanta más @{N}@@{fem}@@{pl}`	`[-cuanta][+cuantas]`
`cuantas más @{adj}@@{fem}@@{pl}`	`[-cuantas][+cuanto]`
`cuanto más @{N}@@{masc}@@{pl}`	`[-cuanto][+cuantos]`
`cuantos más @{adj}@@{masc}@@{pl}`	`[-cuantos][+cuanto]`

A pesar de crear patrones flexibles y de amplia cobertura capaces de tratar el máximo de la casuística posible, para su aplicación es necesario tener en cuenta la ambigüedad que pueden presentar tanto los sustantivos como los adjetivos. Será necesario condicionar, indicándolo en la base de datos, este patrón y aplicarlo solo a aquellos casos en los que estas categorías no tengan un correlato ambiguo; términos como *joven, tonto, ciego* o *azul*, pueden funcionar en el sintagma bien como adjetivo, bien como sustantivo.

Con el fin de evitar la corrección de expresiones correctas como *Cuanto más viejos se hacen, más sabiduría atesoran/*Cuantos más viejos se hacen…* el programa, para los casos en los que sea necesaria una interpretación activa, se limitará a

ofrecer avisos lingüísticos —*posible error de concordancia*— con la información necesaria para llevar a cabo un tratamiento satisfactorio en estas expresiones.

Por otro lado, algunos de estos términos no agotan la ambigüedad en la dicotomía adjetivo-adverbio; en el caso de *medio/media* la ambigüedad se amplía para dar cabida, junto con la versión adjetiva y adverbial, a la versión sustantiva del término. Así, junto al sintagma adjetivo *media hora* podemos encontrar los sintagmas nominales *medio corrupto* o *media rota* y el adverbial *medio mareado.* Para solventar algunos de los errores que surgen en el uso de estas formas, ha sido necesario un estudio preliminar de cada forma con el fin de acotar su ámbito según la categoría que desarrolle. A partir de este estudio de cotextos se han podido desarrollar algunos patrones que con certeza y garantías ofrecen tratamiento a estos errores.

Ante este panorama de ambigüedad múltiple, el programa, una vez más, no ejecutará las correcciones automáticamente. Se limitará a ofrecer avisos al usuario para que, contemplando el caso concreto objeto del tratamiento, y las glosas con la información gramatical, valore y corrija de forma interactiva este tipo de expresiones.

13.1.2 Concordancia intersintagmática; relaciones entre constituyentes oracionales

Otro de los bloques asociados a errores o incidencias relacionadas con la concordancia es este que se ocupa de las relaciones que trascienden los límites del sintagma. Se tratarán, pues, aspectos de las relaciones intersintagmáticas que básicamente se concentran en la relación que establecen el sujeto y el verbo.

Como ya se advirtió, los límites de este trabajo impiden hacer un tratamiento global de las concordancias. No obstante, hay casos conflictivos en los que el usuario presenta dudas o altera la gramática. Estos casos serán tratados particularmente con el fin de dar cobertura a los errores más habituales. Para acometer estos asuntos, se han codificado 137 patrones de error, una recomendación y un aviso lingüístico.

13.1.2.1 *Sustantivos clasificativos*

Se constatan con frecuencia, errores de concordancia entre sujeto y verbo cuando el primero está conformado por estructuras del tipo «*clase/especie/clase* + *de* + N en plural»; **Ese tipo de personas no me agradan*.

Aunque se admiten las dos versiones de concordancia, en singular y en plural, cuando el sujeto presenta como núcleo uno de estos sustantivos clasificativos, —disocian en clases el conjunto al que se aplican— la norma prefiere y recomienda la primera opción, esto es, la concordancia en singular. Habida cuenta de que se trata de una recomendación, y no de un error, estos patrones estarán recogidos y etiquetados en PatErr con una recomendación.

Los sustantivos clasificativos que se han recogido en patrones son los siguientes:

`clase especie gama género tipo suerte variedad`

La misma pauta siguen sustantivos numerales de naturaleza colectiva como:

`par decena docena centena millar` *`etc.`*

que también estarán recogidos bajo la consideración de recomendación. En total se han codificado 28 patrones como el modelo que sigue;

`tipo de @{adj}? @{N}@@{pl} #{vb}##{3ª pl}`	⊗*Ese tipo de buenas personas me inspiran mucho.*
`tipo de @{N}@@{pl} @{adj}? #{vb}##{3ª pl}`	⊗*Ese tipo de personas malvadas me perturban.*

En todos los casos codificados la medida correctiva es la misma; la sustitución del plural en la forma verbal de tercera persona por la variante en singular.

13.1.2.2 *Sustantivos colectivos*

Los nombres colectivos que están insertos en el sujeto de una oración suelen ser términos que dan lugar a errores de concordancia. Sustantivos del tipo *gente, multitud, séquito, alumnado*, etc. —recogidos en la Lista 13—, pese a tener un

significado plural, en su versión singular deben concordar así con el verbo, en singular.

Son comunes expresiones del tipo **Toda esa comunidad tienen los mismos derechos que nosotros, *La familia entera lloraron su muerte*, en las que el escritor establece una concordancia *ad sensum* justificada por la alusión a los individuos que componen el grupo descrito por el sustantivo colectivo. Para evitar estos errores de concordancia se han codificado tres patrones codificados que abarcan la casuística posible:

{Lista 13}@@{sg} @{adj}? #{vb}##{3ªpl}
{Lista 13}@@{sg} @{adv}? @{adj}? #{vb}##{3ªpl}
{Lista 13}@@{sg} @{adv}? @{adj}? que #{vb}##{3ªpl} \w+? #{vb}##{3ªpl}

Como en los casos precedentes, las medidas correctivas se concretan en la sustitución de la forma verbal en plural por la correspondiente en singular.

13.1.2.3 *Pronombres indefinidos*

La discordancia entre sujeto y verbo es habitual cuando el sujeto está desarrollado por alguno de los pronombres indefinidos:

alguno alguna ninguno ninguna

Es común toparse con oraciones agramaticales como ***¿Alguna sabéis cómo hacer un dobladillo?** El error aquí se comete al flexionar en plural la forma verbal. Para evitar estas vacilaciones se han registrado doce patrones que encapsulan todas las posibilidades formales que ofertan estas estructuras.

alguna #{vb}##{2ª pl}
alguna de @{det}+? @{adj}? @{N} @{adj}? #{vb}##{3ª pl}
alguna de @{pn}+ que \w+? #{vb}##{3ª pl}
alguno #{vb}##{2ª pl}
alguno de @{det}+? @{adj}? @{N} @{adj}? #{vb}##{3ª pl}
alguno de @{pn}+ que \w+? #{vb}##{3ª pl}
ninguna #{vb}##{2ª pl}

ninguna de @{det}+? @{adj}? @{N} @{adj}? #{vb}##{3ª pl}
ninguna de @{pn}+ que \w+? #{vb}##{3ª pl}
ninguno #{vb}##{2ª pl}
ninguno de @{det}+? @{adj}? @{N} @{adj}? #{vb}##{3ª pl}
ninguno de @{pn}+ que \w+? #{vb}##{3ª pl}

La medida correctiva en todos los casos será la misma; la sustitución de la forma verbal en plural por la correspondiente en tercera persona de singular.

13.1.2.4 *El verbo faltar*

Otra de las incidencias tratadas en este bloque está representada por algunos usos erróneos del verbo *faltar*, que con frecuencia se trata como un verbo impersonal. El sustantivo o el infinitivo que expresa *lo que falta* será el sujeto de la oración, por lo que es incorrecto mantener el verbo en singular cuando el sustantivo que hace de sujeto es plural; **Falta por venir muchos otros participantes.*

Para subsanar, solo parcialmente, algunos de los casos erróneos que han podido constatarse a partir del uso de este verbo se han registrado cuatro patrones que contemplan algunos de estos contextos. En todos ellos la acción correctiva es la misma; la sustitución de la forma verbal en singular por la correspondiente en plural; **Había faltado tres niños para completar el grupo/Habían faltado tres niños para completar el grupo.*

#faltar##{3ºsg} @{adv}? @{det}? @{pn}@@{pl}
[-#faltar##{3º sg}][+#faltar##{3º pl}]
**Había faltado los vuestros.*

#faltar##{3ºsg} @{adv}? @{det}@@pl? @{adj}? @{N}@@{pl}
[-#faltar##{3º sg}][+#faltar##{3º pl}]
**Falta escasamente tres horas para que venga.*

#faltar##{3º sg} @{adv}? {prep}? {inf} @{det}@@pl+? @{adj}? @{N}@@{pl}
[-#faltar##{3º sg}][+#faltar##{3º pl}]
**Falta por llegar los alumnos de 4º.*

#faltar##{3º sg} @{adv}? {prep}? {inf} @{pn}@@pl
[-#faltar##{3º sg}][+#faltar##{3º pl}]
**Falta todavía por entregar aquellos.*

Para evitar falsos positivos, y debido a la ambigüedad que presenta esta forma, deben condicionarse estos patrones a no ser precedidos ni por un determinante, ni por una preposición —en estos casos *falta* actuaría como sustantivo *la falta/a falta*—, ni por el verbo *hacer* —locución *hacer falta*—.

Como se observa, la identificación precisa de estos errores requiere la codificación de un cotexto amplio codificado que garantice la flexibilidad y validez del patrón.

13.1.2.5 *Estructuras copulativas*

En el uso de expresiones del tipo *ser necesario, ser cierto, ser probable,* es común encontrar errores de concordancia entre el adjetivo que funciona como atributo y el sustantivo o pronombre que desarrolla el sujeto de estas oraciones copulativas; **Era necesario una organización diferente para aumentar el rendimiento; *Es preciso una mayor inversión para estos pueblos; *No es cierto la tesis que defiende.* Para evitar estos errores es necesario hacer concordar en género y número, el adjetivo con el sujeto que suele ir pospuesto.

El diseño de estos patrones está pensado para abordar todas las posibilidades que puedan surgir. Así se han codificado 34 patrones que recogen todas las configuraciones posibles para el sujeto —núcleo con sustantivo o núcleo con pronombre—. En todos los casos la corrección se basará, precisamente, en forzar la concordancia de género y número entre el núcleo del sujeto y el atributo, obedeciendo a la naturaleza del primero.

Se codifican cuatro patrones para un sujeto cuyo sintagma nominal esté desarrollado por un sustantivo, que opcionalmente pueda estar complementado; **Era cierto todas esas graves insidias,* y otros cuatro para la opción de pronombre. Otros adjetivos codificados con el mismo esquema de patrón son: *necesario* y *preciso*.

#ser @{adv}? cierta @{det}+ @{adj}? @{N}@@{masc}
#ser @{adv}? ciertas @{det}+ @{adj}? @{N}@@{masc}
#ser @{adv}? cierto @{det}+ @{adj}? @{N}@@{fem}
#ser @{adv}? ciertos @{det}+ @{adj}? @{N}@@{fem}

#ser @{adv}? cierta @{pn}@@{masc}
#ser @{adv}? ciertas @{pn}@@{masc}
#ser @{adv}? cierto @{pn}@@{fem}
#ser @{adv}? ciertos @{pn}@@{fem}

Para los adjetivos que no ofrecen marcas de género se recogen los siguientes patrones que se replican para las formas *probable* y *posible*.

#ser @{adv}? evidente @{det}+ @{adj}? @{N}@@{pl}
#ser @{adv}? evidentes @{det}+ @{adj}? @{N}@@{sg}

#ser @{adv}? evidente @{pn}@#{pl}
#ser @{adv}? evidentes @{pn}@@{sg}

En todos los casos la corrección se basa en hacer concordar los adjetivos del atributo con el sujeto al que complementan.

13.1.2.6 *Algunas expresiones*

Aunque es más propio del registro oral, no es extraño encontrarse por escrito frases del tipo **Viva los novios*. Las expresiones desiderativas de este tipo cuya estructura básica es «Vb + SN», deben hacer concordar en número el verbo con el sustantivo del sintagma nominal, en el caso expuesto, en plural.

Para acometer los errores generados en estas expresiones se han codificado cuatro patrones que recogen los verbos más invocados en estas estructuras: *vivir* y *morir*. Los patrones base presentan las siguientes estructuras:

viva @{det}@@{pl}+? @{adj}? @{N}
viva @{pn}@@{pl}

Otro caso de discordancia entre el sujeto y el verbo se constata en expresiones en las que está inserta la expresión *hacer público.* Es común leer oraciones incorrectas como **La prensa hizo público ayer las declaraciones del imputado,* en las que se observa la invariabilidad del adjetivo *público* cuando debería establecer concordancia con el sustantivo al que se aplica que forma parte del complemento directo, *declaraciones*.

Para la corrección de este fenómeno se han registrado seis patrones que captan las estructuras posibles en relación con esta expresión. Como en todos los registros de errores, se flexibilizan los patrones para aceptar modificadores o complementos opcionales.

`#hacer público @{det}@@{fem pl}+ @{adj}@@{fem pl}? @{N}@@{fem pl}`
**Hizo público las acusaciones.*

`#hacer público @{det}@@{fem sg}+ @{adj}@@{fem sg}? @{N}@@{fem sg}`
**Hizo público las conversaciones privadas.*

`#hacer público @{det}@@{masc pl}+ @{adj}@@{masc pl}? @{N}@@{masc pl}`
**Hizo público los nombres de los acusados.*

Se contempla, además, la opción de que el complemento directo esté desarrollado por un pronombre:

`la #hacer público`	**la hace público.*
`las #hacer público`	**las hace público.*
`los #hacer público`	**los hace público.*

En todos los casos la medida correctora se basa en hacer concordar el adjetivo *público* con el sustantivo o pronombre al que alude.

Para los casos en los que el programa identifique la expresión ⊗*hacer público* pero no encaje con ninguno de los patrones registrados porque el error no se encuentra en un contexto contiguo, **Hizo público, antes de morir, toda su actividad*

económica, se activará un aviso lingüístico con información sobre el régimen de concordancia de esta expresión.

Por último, se aborda una batería de expresiones similares que adolecen de una falta de concordancia entre constituyentes presentes en oraciones como **Dado la situación presente, renuncio a los beneficios; *Dado la importancia del tema, deberemos asegurarnos antes de firmar.*

En estas estructuras, la gramática exige concordancia en género y número entre el participio y el sustantivo que le sigue. Los siguientes patrones codificados abarcan las posibilidades de concordancia de esta expresión:

`dado @{det}@@{fem sg}+ @{N}@@{fem sg}`	**Dado la crítica recibida*
`dado @{det}@@{fem pl}+ @{N}@@{fem pl}`	**Dado las circunstancias de Amparo*
`dado @{det}@@{masc pl}+ @{N}@@{masc pl}`	**Dado los apoyos recibidos*

13.1.2.7 *Construcciones partitivas*

Uno de los focos generadores de error en relación con las relaciones de concordancia entre el sujeto y el verbo lo constituyen las estructuras partitivas. Estas son especialmente propensas a albergar errores de concordancia debido a que en su interior es fácil que se desdibujen las relaciones entre el sujeto y el verbo; *Es uno de los que más ha ayudado; Es uno de los que más han ayudado.* Para el caso propuesto en el ejemplo, se aceptan las dos variantes; aquella en la que el verbo está en singular —se ha establecido una concordancia *ad sensum*— y la que permite, siguiendo la lógica gramatical, la concordancia en plural; no conviene perder de vista que el sujeto de esta oración es el relativo *que*.

No obstante, no se le ofrece esta libertad de concordancia a las estructuras que presentan primera o segunda persona cuando la construcción forma parte de una oración copulativa cuyo sujeto es un pronombre personal; *yo, tú, usted, vos, nosotros, nosotras, ustedes, vosotros* y *vosotras*. Para estos casos, **Yo soy una de las que quiso votar,* la forma preferida y recomendada por la norma culta es la concordada en plural; *Yo soy una de las que quisieron votar.*

Para subsanar estos errores, junto con otros que surgen a propósito de estas estructuras partitivas, se han codificado varias baterías de patrones que encapsulan todas las posibilidades que pueden darse para cada opción.

- *Primera persona del singular*

yo @{adv}? #ser una? @{adj}? de @{pn} que #{vb}##{1ª sg}
**Yo nunca fui una romántica de esas que seduce con poemas.*

yo @{adv}? #ser uno? @{adj}? de @{pn} que #{vb}##{1ª sg}
**Yo nunca fui un romántico de esos que seduce con poemas.*

La medida correctiva para este repertorio se basa en la sustitución del verbo en la primera persona del singular por la correspondiente forma en tercera persona del plural.

- *Segunda persona del singular*

tú @{adv}? #ser una? @{adj}? de @{pn} que #{vb}##{2ª sg}
** Tú ciertamente eras de esas que estabas atentas.*

tú @{adv}? #ser uno? @{adj}? de @{pn} que #{vb}##{2ª sg}
** Tú ciertamente eras de esos que estabas atentos.*

Este último bloque se ha registrado con todos los posible cotextos para las variantes *vos, usted*. La corrección, en todos los casos, supone la sustitución del verbo en segunda persona del singular por la correspondiente forma en tercera del plural.

13.1.2.8 *Oraciones escindidas*

Algunas de las configuraciones sintácticas más productivas para llevar a cabo la focalización, —entendida aquí como una alteración sintáctica que convierte a un constituyente en el foco de la oración— son las de las oraciones escindidas.

Básicamente el procedimiento consiste en escindir el constituyente que quiere ser focalizado mediante el verbo *ser*. Las estructuras sintácticas resultantes se parecen mucho a las de las oraciones compuestas por subordinadas de relativo; *Fui yo quien compró la cerveza; Quien compró la cerveza fui yo; El que compró la cerveza fui yo; Yo fui quien compró la cerveza.*

En estos ejemplos, se observa que la concordancia que establece el verbo subordinado es precisamente con su sujeto, el pronombre relativo que se infiere con la forma de tercera persona del singular. Lo mismo ocurre con las variantes de segunda persona del singular *tú y vos.* Se excluye *usted* por tener prescriptivamente la concordancia con la forma propia de la tercera persona.

Pero esta operación, en ocasiones, no parece estar automatizada en el hablante y es frecuente que comprometa las normas de concordancia. El error más frecuente establece la relación entre el verbo de la subordinada *compré*, y el sujeto pronominal de la principal *yo*, dando lugar a expresiones poco recomendadas como ⊗*Yo fui quien la compré.*

Si se considera la casuística encerrada en la primera persona de plural, se observa que la pauta cambia; oraciones como *Nosotros fuimos quienes lo dijimos* se consideran aceptables —y recomendadas— frente a estructuras como ⊗*Nosotros fuimos quienes lo dijeron*, que no ofrecen el mismo grado de corrección. Como se observa, aquí la concordancia se establece con el indefinido *quienes,* en lugar de con *nosotros.* Lo mismo sucede en el caso de la segunda persona del plural. Se trata en palabras de Val Alvaró y Mendívil Giró, de una pauta cruzada de concordancia en función del número singular o plural del sujeto pronominal de la oración principal (2011:300).

A la vista de estos datos se puede establecer la siguiente generalización; la concordancia del verbo subordinado en primera y segunda persona del singular se prefiere con el sujeto de relativo antes que con el sujeto pronominal de la principal. Los plurales de la primera y segunda persona, en cambio, establecen la concordancia del verbo subordinado con el sujeto pronominal de la oración principal, y no con el sujeto de la subordinada.

Para dar solución global a este fenómeno de *pauta cruzada,* se han codificado 56 patrones en el PatErr.

A continuación, se ofrece el repertorio que encapsula todas las combinaciones posibles en cuanto a género y estructura de los constituyentes para la primera persona del singular.

yo @{adv}? #ser quien @{pn}? #{vb}##{1ª sg}	[-#{vb}##{1ª sg}] [+#{vb}##{3ª sg}]
yo @{adv}? #ser la que @{pn}? #{vb}##{1ª sg}	[-#{vb}##{1ª sg}] [+#{vb}##{3ª sg}]
yo @{adv}? #ser el que @{pn}? #{vb}##{1ª sg}	[-#{vb}##{1ª sg}] [+#{vb}##{3ª sg}]
quien @{adv}? @{pn}? #{vb}##{1ª sg} #ser yo	[-#{vb}##{1ª sg}] [+#{vb}##{3ª sg}]
la que @{adv}? @{pn}? #{vb}##{1ª sg} #ser yo	[-#{vb}##{1ª sg}] [+#{vb}##{3ª sg}]
el que @{adv}? @{pn}? #{vb}##{1ª sg} #ser yo	[-#{vb}##{1ª sg}] [+#{vb}##{3ª sg}]
#ser yo quien @{adv}? @{pn}? #{vb}##{1ª sg}	[-#{vb}##{1ª sg}] [+#{vb}##{3ª sg}]
#ser yo la que @{adv}? @{pn}? #{vb}##{1ª sg}	[-#{vb}##{1ª sg}] [+#{vb}##{3ª sg}]
#ser yo el que @{adv}? @{pn}? #{vb}##{1ª sg}	[-#{vb}##{1ª sg}] [+#{vb}##{3ª sg}]

La solución para las configuraciones erróneas codificadas y sus derivadas, **Yo no fui quien lo tiré a la basura; *Quien siempre perdono soy yo; *La que no voy soy yo,* se basa en la sustitución del verbo en primera persona del singular por la correspondiente en tercera persona. Como se ha dicho, esta batería se aplica igualmente a las variantes de segunda persona, *tú* y *vos*.

Los patrones se han pensado, en cuanto al diseño, para tener la mayor flexibilidad posible dentro de los límites de los contextos que se pueden abordar con el recurso que aquí se presenta. Podrán capturarse, así, secuencias complejas como **Yo no fui la que le tiré de la lengua,* que abarcan una gran parte de la casuística de estas estructuras independientemente del tiempo del verbo *ser,* del verbo subordinado y de la inclusión de adverbios o pronombres en la expresión.

Como se observará, se codifican desglosadas todas las variantes de género y persona para facilitar tanto el tratamiento automático como la legibilidad del patrón.

A continuación, se detalla el repertorio recogido para las dos variantes de la primera persona del plural, *nosotros* y *nosotras*:

#ser nosotras las que @{adv}? @{pn}? #{vb}##{3ª pl}	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
#ser nosotras quienes @{adv}? @{pn}? #{vb}##{3ª pl}	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
#ser nosotros los que @{adv}? @{pn}? #{vb}##{3ª pl}	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
#ser nosotros quienes @{adv}? @{pn}? #{vb}##{3ª pl}	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
las que @{adv}? @{pn}? #{vb}##{3ª pl} #ser nosotras	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
los que @{adv}? @{pn}? #{vb}##{3ª pl} #ser nosotros	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
nosotras @{adv}? @{pn}? #ser las que #{vb}##{3ª pl}	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
nosotras @{adv}? @{pn}? #ser quienes @{pn}@@{reflex}? #{vb}##{3ª pl}	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
nosotros @{adv}? @{pn}? #ser los que #{vb}##{3ª pl}	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
nosotros @{adv}? @{pn}? #ser quienes @{pn}@@{reflex}? #{vb}##{3ª pl}	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
quien @{adv}? @{pn}? #{vb}##{3ª pl} #ser nosotras	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]
quien @{adv}? @{pn}? #{vb}##{3ª pl} #ser nosotros	[-#{vb}##{3ª pl}] [+#{vb}##{1ª pl}]

La corrección, en este caso, se basa en la supresión de la tercera persona del plural por la que le corresponde, la primera persona del plural.

Están registrados, además, los patrones correspondientes para la segunda persona de plural *vosotros, vosotras* que, en lo estructural, son una réplica del expuesto. La corrección aquí eliminará el verbo en tercera persona e incluirá la forma verbal correcta en segunda persona.

Tras una tarea contrastiva aplicada sobre otros correctores, se ha observado que estos casos de vacilación de concordancias y pauta cruzada en estructuras escindidas pasan desapercibidos para estos sistemas de revisión.

13.1.3 Concordancias de género

Con relación a la concordancia que debe establecerse entre determinantes, sustantivos y adjetivos por cuestiones de género, se han rescatado algunos de los casos que, por sus particularidades, con frecuencia encierran dificultad o son causa de discordancia[94].

Para acometer este grupo de incidencias, se han registrado 85 patrones de error y una recomendación que identifica sustantivos y adjetivos que, aunque tienen posibilidad de flexionar y concordar con los dos géneros, solo uno es el recomendado por la norma —*mimbre, polígloto,* etc.—.

13.1.3.1 *Vacilación en el género de algunos sustantivos*

Se ha localizado un grupo de sustantivos masculinos que resultan confusos en cuanto al género por su terminación opaca —*alambre*— o por su terminación equívoca en -*a,* —*aneurisma*—*.*

Todo este repertorio está etiquetado en el DRAE como sustantivos masculinos, aunque por su terminación en cuanto a la información de género, es habitual verlos concordados erróneamente en femenino. Los sustantivos masculinos registrados y codificados son los que siguen:

alambre	aceite	avestruz
aneurisma	pijama	interrogante
apéndice	vinagre	mimbre
apocalipsis	detonante	acné
arroz	contraluz	tequila
áspid	tanga	vodka

El patrón general para este grupo de sustantivos masculinos será como el que sigue:

```
@{det}@@{fem}+ @{adj}? @alambre
```

[94] Para facilitar el acceso a los datos, los patrones que corrigen la concordancia de género, recibirán la etiqueta 504 mientras que aquellos que tratan la flexión de número se recogen en la incidencia 506.

La corrección se basará en la incorporación de la flexión en femenino de todos los complementos que le bordeen, tanto determinantes como adjetivos.

Complementariamente, se registran los casos contrarios, que tratan con sustantivos femeninos que erróneamente suelen establecer concordancias en masculino. Los sustantivos femeninos que se han recogido son los siguientes:

apócope	comezón	porción
cochambre	hinchazón	parálisis
pelambre		

A partir del siguiente patrón general, han sido codificados todos estos sustantivos. La corrección, como en el caso anterior, se basará en la incorporación de la flexión femenina en todos los complementos contiguos que le afecten.

```
@{det}@@{masc}+ @{adj}? @parálisis
```

Se ha encontrado, dentro de este ámbito, un grupo reducido de sustantivos vacilantes para los que la norma expone sus recomendaciones en cuestión de género sin llegar a prescribir su concordancia. Para estos casos, se ha previsto una recomendación que recoge términos que pueden desarrollar ambas flexiones, en ocasiones, dependiendo del sentido al que se pretenda hacer referencia. Se han recogido los siguientes sustantivos para los que la norma recomienda la concordancia con sus complementos en masculino:

macro	mimbre	margen
dote	tizne	vislumbre

13.1.3.2 *Profesiones u oficios*

Cuando el nombre de una profesión es un compuesto formado por un sustantivo y un adjetivo que lo acota, es necesaria la concordancia de género entre ellos. La regla, como es lógico, se expande al resto de los determinantes —si los hubiera—, que constituyen el sintagma nominal.

Se codifican, a este respecto, tres patrones que recogen los casos que con más frecuencia generan dudas en el hablante;

@{det}@@{fem}+ @{adj}? @detective @privado@@{masc}
**La detective privado*

@{det}@@{fem}+ @{adj}? @guarda @jurado@@{masc}
**La guarda jurado*

@{det}@@{fem}+ @{adj}? @primer@@{masc} @ministro@@{masc}	
**La primer ministra*	**La primera ministro*

Puede ser útil, como trabajo futuro, ampliar este elenco a partir de un estudio de las colocaciones que involucran a sustantivos que dan nombre a profesiones. Este análisis permitirá extraer los compuestos más frecuentes de nuestra lengua.

13.1.3.3 *Cuantificadores e indefinidos*

Es común leer expresiones como **Fue una de los mejores ministros*, en las que la concordancia no se lleva a término como es de esperar. Para este tipo de construcciones partitivas en las que el cuantificador presenta flexión —*uno/a, mucho/a, ninguno/a,* etc.— es necesario establecer concordancia de género y número con el sustantivo que aparece en el complemento desarrollado por un sintagma preposicional, en este caso *ministros.* Por otro lado, no es correcto utilizar el cuantificador o indefinido en femenino si el grupo al que se alude, *ministros,* está compuesto de individuos de ambos sexos.

Para subsanar estas discordancias, se codifican pares de patrones de error como el que sigue:

`alguna de los`	`[-alguna de los]` `[+alguno de los]`	**Alguna de los que han asistido*
`algunas de los`	`[+algunas de los]` `[+algunos de los]`	**Algunas de los que han asistido*

Se replican estos patrones para dar cabida a los siguientes pronombres:

`@mucha` `@ninguna` `@otra` `@una` `varias`

13.1.3.4 *Algunos sintagmas preposicionales*

Es común encontrar errores de concordancia como **El Estado solo obedece a razones de tipo económicas; *Ofrecen ayudas de carácter sociales; *Esgrimió motivos de índole sanitario*, en construcciones preposicionales que sirven de complemento a ciertos sustantivos.

Locuciones similares a *de tipo, de carácter* o *de corte,* que suelen ir seguidas de un adjetivo, deben presentar concordancia en singular y en el género del sustantivo que sigue a la preposición; *tipo, carácter, índole, estilo*, etc.

Para subsanar estos errores, se han codificado cuatro pares de patrones que recogen los sustantivos femeninos más utilizados en estas estructuras;

índole naturaleza motivación razón

de índole @{adj}@@{masc}	[-@{adj}@@{masc}][+@{adj}@@{fem}@@{sg}]
de índole @{adj}@@{pl}	[-@{adj}@@{pl}][+@{adj}@@{sg}@@{fem}]

Por su parte, se incluyen tres pares de patrones para tratar los casos en los que se presentan sustantivos masculinos: *estilo, carácter* y *tipo*.

de estilo @{adj}@@{fem}	[-@{adj}@@{fem}][+@{adj}@@{masc}@@{sg}]
de estilo @{adj}@@{pl}	[-@{adj}@@{pl}][+@{adj}@@{sg}@@{masc}]

En todos estos patrones se veta la forma en plural del adjetivo y se fuerza la concordancia según sea el género del sustantivo al que se refiere.

13.1.3.5 *Tratamientos de respeto*

Es común, por la confusión que generan los conceptos de género y sexo, encontrar expresiones incorrectas como **Su despistado señoría; *El Majestad vino a saludarnos.* Los sustantivos que, como los del ejemplo, muestran tratamientos de respeto,

alteza majestad señoría excelencia

llevan sus complementos inmediatos —determinantes y adjetivos— en femenino, de acuerdo con el género de estos sustantivos. No se debe, en estos casos, hacer concordar el sustantivo con el sexo del referente.

No obstante, si el adjetivo no está en posición contigua —función de atributo o de predicativo; *Su señoría estaba desnortado; Su majestad llegó exhausto*— debe presentarse en el género que corresponda al sexo del referente. Lo mismo sucede con otros elementos no adyacentes como los pronombres; *Su Alteza* (Felipe VI) *era el que no podía venir.*

Para solventar los errores que puedan surgir en contextos de contigüidad se ha registrado el siguiente patrón para cada uno de estos sustantivos;

```
@{det}@@{masc} @{adj}@@{masc}? @excelencia @{adj}@@{masc}
```

13.1.3.6 *Género de algunos numerales*

El ámbito de algunos numerales, tanto en su versión ordinal como en la cardinal, provoca, con frecuencia, dudas y errores de concordancia en relación con el género.

Por un lado, los adjetivos determinativos ordinales *primer* y sus derivados, así como *tercer*, y los suyos, deben concordar en género con los sustantivos femeninos a los que identifiquen; **Fue su primera actuación en aquella ciudad; *Es la decimotercera vez que te hago la misma pregunta*. Es erróneo, por lo tanto, emplear la versión apocopada cuando el sustantivo que le sigue es femenino.

`@{det}@@{fem} primer`	**La primer novia que tuvo fue la mejor para él.*
`primer @{N}@@{fem}`	**Vente a primer hora.*

Este patrón, que se duplica para el adjetivo *tercer*, tiene marcado afirmativamente, en su línea de registros, el campo de relaciones morfoléxicas. De esta manera el patrón no solo dará cobertura a las secuencias codificadas en el patrón —*primer*—, sino a todas las formas compuestas en las que aparezca; *vigésimo primer, trigésimo primer,* etc. Para todos estos casos la medida correctiva será la misma; la sustitución del numeral apocopado por la versión en femenino.

En relación con el ámbito de sustantivos cardinales, es común encontrar expresiones discordantes como **las miles de mujeres; *las cientos de alumnas* o **las millones de pesetas.* Pero *mil, ciento o millón* como sustantivos, son masculinos, y por lo tanto el artículo que les precede debe ser masculino para entablar concordancia. Así, el programa, corregirá estos casos modificando el género del determinante; *los miles de mujeres; los cientos de alumnas; los millones de pesetas.*

El patrón registrado para cada sustantivo listado es como el que sigue:

ciento mil millón billón trillón

@{det}@@{fem}? cientos de	[-@{det}@@{fem}][+@{det}@@{masc}]

Como se observa, la corrección se basa en la sustitución del determinante o los determinantes femeninos que preceden al término numeral, por los correspondientes en masculino.

13.1.3.7 *Excepciones; algunos sustantivos femeninos*

Hay un grupo cerrado de sustantivos que en ciertos contextos presentan un régimen especial en relación con la concordancia. Este grupo, congregado en la Lista 1, está formado por los sustantivos femeninos que comienzan con *á-; há-* tónica y exigen cuando están en singular, ir precedidos por la forma masculina del determinante; *El águila tenía un hambre canina;*

Esta regla general presenta ciertas restricciones. En primer lugar, su aplicación solo afecta a cuatro determinantes; los artículos en singular, *el* y *un* y los adjetivos determinativos indefinidos que contienen esta última forma; *algún* y *ningún.* Para estos casos propios de la regla general se registra el siguiente patrón global:

@@det@@masc^({art}^@algún^@ningún) {Lista 1 sing}
**esa área, *aquellos áreas, *tanto área, *primera área*

Y dos baterías de patrones que recogen explícitamente el error, y facilitan el procesamiento;

`la {Lista 1 sing}`	**el área*
`una {Lista 1 sing}`	**un área*
`alguna {Lista 1 sing}`	**alguna área*
`ninguna {Lista 1 sing}`	**ninguna área*

`los {Lista 1 pl}`	**los áreas*
`unos {Lista 1 pl}`	**unos áreas*
`algunos {Lista 1 pl}`	**algunos áreas*
`ningunos {Lista 1 pl}`	**ningunos áreas*

Por otro lado, el determinante y el sustantivo tienen que ocupar posiciones contiguas; si entre ellos se interpone cualquier elemento, el determinante retoma su forma femenina; *El asesino utilizó la misma arma en los dos sucesos; Nunca digas de esta clara agua no beberé; Se reunieron en la inmensa área metropolitana*. A pesar de estas concordancias especiales, este grupo de sustantivos son de género femenino a todos los efectos; cualquier elemento con el que mantenga concordancia lo hará con flexión femenina.

Para el tratamiento de estos casos en los que se intercala un adjetivo, es decir, se rompe la contigüidad entre determinante y sustantivo se han codificado las siguientes baterías de patrones:

`el @{adj}@@{masc} {Lista 1 sing}`	**el gran área*
`un @{adj}@@{masc} {Lista 1 sing}`	**un impresionante área*
`algún @{adj}@@{masc} {Lista 1 sing}`	**algún inmenso área*
`ningún @{adj}@@{masc} {Lista 1 sing}`	**ningún pequeña área*

`los @{adj}@@{masc} {Lista 1 sing}`	**los grandes áreas*
`unos @{adj}@@{masc} {Lista 1 sing}`	**unos impresionantes áreas*
`algunos @{adj}@@{masc} {Lista 1 sing}`	**algunos inmensos áreas*
`ningunos @{adj}@@{masc} {Lista 1 sing}`	**ningunos pequeñas áreas*

La regla general, aparte de las restricciones de aplicación expuestas, tiene sus excepciones. Se retomará el artículo femenino con:

I. los nombres de letras: *la hache*; *la alfa*; *la a*.
II. los nombres propios de mujer —en aquellos contextos en que pueden aparecer con artículo—, *La Amparo que tú decías era una descarada*; *Esta no es la Ana que yo conocía*[95].
III. los sustantivos generados por las siglas: *la AMPA*; *la AECC*.
IV. los topónimos: la Ávila teresiana; *la Asia Central del siglo XX*.

Por último, cabe señalar que esta regla se invalida cuando el sustantivo femenino pierde tonicidad. Tal es el caso de **una amita de casa, *la alita de pollo, *la habita de la suerte*, que al componerse con un diminutivo pierden la tonicidad de la primera sílaba.

Todas estas consideraciones han quedado reflejadas en la Lista 1, que se ha elaborado a partir de la extracción del lexicón mediante consultas sobre todos los sustantivos femeninos que comienzan con *á-, há-.* De este primer repertorio, de nuevo mediante consultas y apoyándonos en nuestro etiquetario, se han retirado los nombres propios, —de persona y topónimos—, las siglas y los nombres de las letras. Queda así conformada la Lista 1 que operará en estos patrones.

13.1.4 Concordancias de número

Como ha podido observarse, muchos de los errores que con frecuencia surgen a propósito de la concordancia de número han sido recogidos ya en otros bloques, especialmente en el de concordancias intersintagmáticas en relación con los vínculos y coincidencias que se establecen entre el sujeto y el verbo de una oración.

Habida cuenta del amplio grado de granularidad que presenta el diseño y la tipología de errores de aquí se propone, el hecho de que haya solapamientos y

[95] El uso en estos contextos es fluctuante; el mismo nombre puede comportarse de manera diferente dependiendo de si es nombre de mujer o no: *El África subsahariana está por descubrir; La África con la que hablabas no es la misma que yo conozco.*
Para estos casos de homófonos se seguirá este criterio; el programa aplicará el determinante femenino para el nombre de mujer, y el masculino para los topónimos.

subsunciones en algunos fenómenos, o de que exista la posibilidad de adscribir un mismo patrón a diferentes incidencias, es habitual en muchos de los registros de PatErr.

Junto con los patrones relacionados con el número expuestos en epígrafes anteriores, se han recogido otros cuantos casos residuales de términos o expresiones que generan duda o error. Se listan, a continuación, los temas tratados:

1. Términos compuestos como *rodapié* o *traspié* son frecuentemente tomados como una forma invariable de apariencia plural —*crisis*—. El caso de estas palabras supone una excepción con respecto a otros compuestos formados por *-pié —ciempiés, buscapiés, reposapiés*, etc.— que son invariables en plural. Se han codificado sendos patrones modificando la flexión del sustantivo cuando se encuentra en un contexto singular. Con ellos se corregirán errores como: **Se dio un traspiés que casi se mata; Finalmente elegí el mejor rodapiés de madera.*

```
@{det}@@{sg}+ @{adj}? traspiés
@{det}@@{sg}+ @{adj}? rodapiés
```

2. Algunas expresiones porcentuales son también focos de discordancia cuando funcionan como sujeto de una oración copulativa; **Casi el 90% de las propuestas fue extravagante.* La concordancia con el verbo en estos casos debe establecerse en plural, como el sustantivo de la expresión del porcentaje; *El 60% de las decisiones son beneficiosas.* Para atrapar todas las posibilidades que presenta esta restricción de las expresiones porcentuales, se han codificado los siguientes patrones:

`El \d+ % de los @{adv}? @{N} #ser##{3ª sg}`	`[-#ser##{3ª sg}]` `[+#ser##{3ª pl]`
`El \d+ % de las @{adv}? @{N} #ser##{3ª sg}`	`[-#ser##{3ª sg}]` `[+#ser##{3ª pl]`
`El @{numeral}@@{card} por ciento de las @{N} @{adv}? #ser##{3ª sg}`	`[-#ser##{3ª sg}]` `[+#ser##{3ª pl]`
`El @{numeral}@@{card} por ciento de los @{N} @{adv}? #ser##{3ª sg}`	`[-#ser##{3ª sg}]` `[+#ser##{3ª pl]`

3. Por último, se han codificado algunas expresiones en las que interviene el verbo *dar*, que con frecuencia presentan errores de concordancia en relación al número que adopta el verbo.

#dar##{3ª sg} ganas de	[-#dar##{3ª sg}] [+#dar##{3ª pl}]
#dar##{3ª sg} escalofríos de	[-#dar##{3ª sg}] [+#dar##{3ª pl}]

En ambos ejemplos los verbos deben ir en tercera persona del plural en coincidencia con el sustantivo de la expresión, *Daban ganas de pedirle una cita; Me dan escalofríos de pensarlo.*

13.1.5 Concordancias de persona

Como en el epígrafe dedicado al número, se recogen aquí casos residuales que tratan sobre los errores de concordancia en cuanto a la persona gramatical y no han sido abordados en líneas anteriores.

Un error frecuente dentro de este ámbito de correspondencias entre personas gramaticales es el que surge en estructuras del tipo *dar más de sí.* Es común encontrar oraciones como **No doy más de sí; *Ayer estuviste fuera de sí.* Estas oraciones no son correctas porque no respetan la concordancia entre la persona del sujeto y la que se adjudica al reflexivo. Según las reglas de nuestra gramática esta correspondencia debe establecerse del siguiente modo;

1ª pers. sg. ***yo***	mí
2ª pers. sg. ***tú***, ***vos***	ti
3ª pers. sg. y pl. ***él***, ***ella***, ***ellos***, ***ellas***	sí

1ª pers. pl. **nosotros, nosotras**	sí
2ª pers. pl. **vosotros, vosotras**	sí
2ª pers. **usted, ustedes**	sí

Respetando estas pautas de concordancia, se han codificado cuatro pares de patrones análogos al proponemos para dar cobertura a otras expresiones:

`dar más de sí` `estar fuera de sí` `volver en sí` `caber en sí`

`#dar##{1ª sg} más de sí`
`#dar##{2ª sg} más de sí`

Para el caso del último patrón de segunda persona, es necesario poner como condición que a la secuencia identificada como error no le preceda *usted; Usted no cabe en sí de gozo.*

En ambos patrones el tratamiento será el mismo; la sustitución del pronombre reflexivo incorrecto, *sí*, por el correspondiente; *mí* o *ti*.

Por otro lado, y a propósito de estas correlaciones que se establecen entre las personas gramaticales, se codifica una batería de patrones para corregir los usos discordantes más frecuentes de las formas *usted* y *ustedes* que se detectan tanto en el verbo como en los pronombres que le pueden circundar.

`usted @{adv}? #{vb}##{2º sg}`	`[-#{vb}##{2º pers sg}]` `[+#{vb}##{3º pers sg}]`
**Usted nunca vienes.*	

`usted @{adv}? te`	`[- te][+ se]`
**Usted no te calles.*	

`te #{vb}##{2ª sg} usted`	`[-te #{vb}##{2ª pers sg}]` `[+se #{vb}##{3ª pers sg}]`
**Te caes usted.*	

a usted @{adv}? te	[- te][+ le]
**A usted te gustan los cotilleos.*	

ustedes @{adv}? #{vb}##{2º pl}	[-#{vb}##{2º pers pl}] [+#{vb}##{3º pers pl}]
**Ustedes casi no coméis.*	

ustedes @{adv}? os #{vb}##{2ª pl}	[-os #{vb}##{2ª pers pl}] [+se #{vb}##{3ª pers pl}]
**Ustedes os venís.*	

a ustedes @{adv}? os	[-te #{vb}##{2ª pers sg}] [+se #{vb}##{3ª pers sg}]
**Te caes usted.*	

os #{vb}##{2ª pl} ustedes	[-os #{vb}##{2ª pers pl}] [+se #{vb}##{3ª pers pl}]
**A ustedes no os permito venir.*	

Finalmente, se ha codificado un patrón que corrige oraciones del tipo **Vosotras se os volvéis a casa; Se os laváis los dientes antes de ir a la cama.* A pesar de ser un error más frecuente en la lengua hablada, es un caso extendido en algunas zonas meridionales de España.

se os #{vb}##{2º pl}	[- se][+ os]	**Vosotras se os volvéis a casa.*

13.2 Gramática verbal

Habida cuenta de los límites y limitaciones de este trabajo, y de las técnicas empleadas para el desarrollo y explotación de este recurso —bajo nivel—, conviene insistir en que una buena parte de los casos erróneos por causas sintácticas que suelen necesitar un análisis profundo de toda la secuencia no encontrarán tratamiento en el repositorio de PatErr.

Por otro lado, hay otros errores relacionados con el régimen verbal, como los usos impropios del gerundio, que necesitarán algún tipo de análisis semántico para poder certificar que estos usos son indebidos. El tratamiento de estos casos, igualmente, deberá ser aplazado hasta poder integrar algún recurso de PLN capaz de llevar a cabo una interpretación semántica.

A pesar de las limitaciones impuestas por los recursos y planteamientos que aquí se adoptan para llevar a cabo la revisión textual, se han registrado 151 patrones de casos inscritos dentro de los límites de la gramática verbal que permiten asesorar y subsanar errores frecuentes de la gramática de nuestra lengua en los que el verbo es el término que genera el error. Una gran parte de estos errores, que comprometen la gramática de ciertas expresiones, pasan desapercibidos para los correctores que han sido examinados.

13.2.1 Usos del gerundio

En términos generales puede definirse el gerundio como una forma verbal impersonal que expresa simultaneidad o anterioridad de la acción con respecto al tiempo en que se inscribe el texto. Se requiere, por tanto, un análisis profundo de los constituyentes, tanto sintáctico, —reconocimiento del sujeto—, como semántico, para poder establecer un tratamiento global que recoja todos sus usos erróneos.

No obstante, con los recursos disponibles se ha podido codificar uno de sus usos no recomendable protagonizado por gerundios semilexicalizados muy frecuentes; aquellos que expresan relación de exclusión o inclusión en un grupo.

Es común, partiendo de estas formas, generar expresiones no recomendadas en las que el gerundio se adhiere directamente al complemento de persona al que se refiere; ⊗*Los clientes, incluyendo los niños, tendrán que presentar su documentación identificativa*. Aunque la norma no lo considera un uso erróneo, sí desaconseja este tipo de estructuras en las que se omite la preposición *a.*

Los patrones que se han codificado para tratar estos usos desaconsejados ofrecen todas las posibilidades estructurales de un sintagma nominal que

típicamente sigue a estos gerundios. Las configuraciones que se han registrado son las siguientes:

incluyendo @{N}@@{prop}	[-incluyendo][+incluyendo a]
excluyendo @{N}@@{prop}	[-excluyendo][+excluyendo a]
exceptuando @{N}@@{prop}	[-exceptuando][+exceptuando a]

incluyendo @{det}+? @{adj}? @{N}	[-incluyendo][+incluyendo a]
excluyendo @{det}+? @{adj}? @{N}	[-excluyendo][+excluyendo a]
exceptuando @{det}+? @{adj}? @{N}	[-exceptuando][+exceptuando a]

La medida correctiva en todos los casos es la misma; la adición de la preposición que rige el complemento de persona.

La otra posibilidad que ofrece el sintagma nominal es la desarrollada por un pronombre; ⊗*Podéis venir todos, incluyendo tú*. Para estos casos la opción más recomendable que propone la norma es sustituir el gerundio por el participio correspondiente; *Podéis venir todos, incluido tú*.

Para el tratamiento de este caso ha sido necesaria la codificación manual de todas las posibilidades que puedan desarrollarse, debido a que cada una presenta una corrección propia que requiere la concordancia del participio con el pronombre personal que forme parte del patrón primario.

A continuación, se presentan los patrones dedicados a la forma *incluyendo* junto con su acción correctiva. Como se observa, se incluyen tanto la opción voseante como las formas de tratamiento de respeto; *usted* y *ustedes.*

incluyendo yo	incluido yo
incluyendo tú	incluido tú
incluyendo él	incluido él
incluyendo ella	incluida ella
incluyendo nosotros	incluidos nosotros
incluyendo nosotras	incluidas nosotras
incluyendo vosotros	incluidos vosotros
incluyendo vosotras	incluidas vosotras

incluyendo ellos	incluidos ellos
incluyendo ellas	incluidas ellas
incluyendo vos	incluido vos
incluyendo ustedes	incluidos ustedes
incluyendo usted	incluidos usted

Esta misma batería se ha desarrollado con los otros dos gerundios susceptibles de formar estas estructuras: *excluyendo* y *exceptuando.*

13.2.2 Usos del infinitivo

Uno de los errores más comunes que surgen con el uso de los infinitivos es aquel en el que se emplea el infinitivo para introducir información dirigida a alguien; **Antes de nada, decir que es un placer veros a todos aquí.*

Este tipo de estructuras, propias del registro oral periodístico, se están extendiendo a la lengua escrita cualquiera que sea el registro o tipo de texto, y constituyen un error de base debido a que nuestra gramática exige como principio básico de construcción oracional que las oraciones presenten un verbo conjugado. Para evitar o subsanar estos casos, la norma culta recomienda sustituir el infinitivo por una forma flexionada o una perífrasis; *Antes de nada, me gustaría decir…; Antes de nada, deciros que…*

Como puede intuirse, los recursos disponibles y el planteamiento que se ha propuesto para llevar a cabo la corrección no permiten ofrecer una corrección precisa para estos casos en los que el tratamiento no puede acotarse y está sujeto al contexto comunicativo en el que se inscriba la secuencia.

Por este motivo se han codificado como avisos lingüísticos las secuencias erróneas más frecuentes relacionadas con este fenómeno:

```
En primer lugar, decir que
En segundo lugar, decir que
En tercer lugar, decir que
Por último, decir que
```

para las que se ha previsto la siguiente glosa que ayudará al escritor a subsanar estas secuencias:

> Este tipo de expresiones características de la lengua oral, no se consideran recomendables por la norma culta ni tampoco para textos escritos. Se prefiere sustituir el infinitivo (que provoca el error), por una forma verbal flexionada, una perífrasis o una construcción análoga. La oración ⊗*En primer lugar, señalar que este proyecto es fruto del trabajo conjunto de ambas naciones,* puede sustituirse por: *En primer lugar, hemos de señalar/quisiera señalar/conviene señalar,* etc.

Estos avisos lingüísticos han sido desarrollados para otros verbos que suelen participar de estas estructuras:

informar comentar indicar señalar

Otro conjunto de patrones que ha sido registrado bajo la etiqueta de aviso lingüístico en relación con esta forma no personal de los verbos tiene por objeto el uso inconveniente del infinitivo con valor exhortativo sustituyendo la forma en imperativo o en subjuntivo, que son las que corresponden para estos enunciados. Estas secuencias, tan habituales en la lengua oral, están especialmente desaconsejadas para los textos escritos.

Como en el caso precedente, el programa no puede ofrecer una opción correcta e inequívoca para cada patrón debido a que no cuenta con la información contextual necesaria para ofrecer una solución satisfactoria —intención comunicativa, interlocutores, etc.—. Por este motivo el usuario contará con una glosa y ejemplos de uso que le permitirán corregir su texto —de forma interactiva— conforme a los usos propuestos por la norma.

A continuación, se ofrecen los patrones registrados para estos casos, junto con un ejemplo de uso erróneo que los han motivado:

No? {inf} @{det}+? @{adj}? @{N}	**Resolver la ecuación siguiente.*
No? {inf} {prep}	**No venir por las tardes.*
No? {inf} #{ger}	**Ir corriendo.*
No? {inf} @{adv}	** No comer deprisa.*
No? {inf} @{pn}@@{sujeto}	**Comprar vosotros.*

Las glosas que se han previsto para ofrecerse al usuario dependerán de si el patrón identificado presenta una partícula de polaridad negativa. Según sea uno u otro caso, la recomendación para subsanar el error difiere.

Para los casos que presentan negación en el patrón de error, la información será la siguiente:

> El infinitivo, que a menudo se utiliza en fórmulas de sentido exhortativo para solicitar algo, no es correcto para este tipo de órdenes negativas en las que debe utilizarse el presente de subjuntivo. Así, la expresión más recomendada para peticiones negativas del tipo **No hacer caso*, sería *No haga/hagas caso*. No obstante, el uso del infinitivo se acepta con más naturalidad en casos en los que el interlocutor no es específico; *No fumar*.

Si el patrón identificado se presenta en positivo, la glosa se desarrollará de otro modo:

> El infinitivo, que a menudo se utiliza en fórmulas de sentido exhortativo para solicitar algo, no está recomendado en este tipo de expresiones, para las que la norma prefiere un uso del imperativo. Así, la expresión más recomendada para peticiones del tipo **Hacer un esquema*, sería *Haz/ haced un esquema*. El uso del infinitivo se acepta con más naturalidad en los casos en los que el interlocutor no es específico; *No fumar*.

13.2.3 Vacilación entre tiempos o modos verbales

Los fenómenos tratados bajo este epígrafe tienen relación, bien con errores derivados de las limitaciones y restricciones que imponen algunos verbos, bien con estructuras inadecuadas relacionadas con el régimen verbal.

Se tratarán, adscritos a una etiqueta de error que ofrecerá tratamiento automático, fenómenos vinculados a la impersonalidad, más precisamente aquellos que presentan el uso del verbo impersonal *haber* obviando la naturaleza de este; **Habrán muchas sorpresas en la fiesta de Luis*.

Por otro lado, se tratarán asuntos más generales relacionados con la vacilación entre tiempos o modos verbales.

13.2.3.1 *Usos no impersonales de verbos que lo son*

El verbo haber

Es frecuente y creciente la proliferación de oraciones como **Habían muchas personas en el congreso*. Estos usos ya extendidos tanto en el discurso oral, como en el escrito tienen su mayor difusión en algunos países de América y en la zona oriental de España.

El verbo *haber*, además de emplearse como verbo auxiliar, tiene un valor existencial en su versión impersonal, es decir, en tercera persona del singular. Las formas correctas para esta expresión de la existencia de algo mediante el verbo *haber* serán, pues, las propias de los verbos con estas restricciones; infinitivo, gerundio, participio y las formas terciopersonales en singular: *hay, había, hubo, habrá, habría, haya,* etc.

El germen de este error parte de obviar estas restricciones e interpretar que el sintagma nominal que sigue a la forma verbal actúa como sujeto, en lugar de como complemento directo, por lo que es errónea la concordancia que el hablante pretende establecer entre el verbo y el grupo nominal.

Los patrones que se han codificado para la detección de estos errores tomarán las formas plurales de tercera persona, y la corregirán por la forma en singular, esto es, por la forma impersonal. Junto a la forma, y para acotar el error con precisión y no confundirlo con alguno de los usos legítimos de este verbo, es necesario desarrollar parte del cotexto que rodeará a la forma, tanto en el margen derecho como en el izquierdo. En el caso del primero, siempre será desarrollado por un sintagma nominal.

Se codifican, asimismo, otras baterías de patrones que recogen las perífrasis verbales compuestas por el verbo *haber*, «*ir + haber + a*», «*poder + haber*» que igualmente tendrán la restricción de construirse exclusivamente con la tercera persona del singular del verbo que le preceda. A partir de estos patrones, podrán tratarse y corregirse errores como: **Van a haber unos días soleados; *La sala estaba*

llena; podrían haber unas mil personas, errores que pasan desapercibidos para el corrector de *Word*.

Se ofrecen a continuación, las baterías de patrones desglosadas que se han codificado para corregir estos usos incorrectos. Su diseño globalizador obedece a la intención de proporcionar cobertura a toda la casuística que se relacione con este fenómeno.

han @{det}+ @{adj}? @{N}	[-han][+hay]
habrán @{det}+ @{adj}? @{N}	[-habrán][+habrá]
hayan @{det}+ @{adj}? @{N}	[-hayan][+haya]
habían @{det}+ @{adj}? @{N}	[-habían][+había]
habrían @{det}+ @{adj}? @{N}	[-habrían][+habría]
hubieran @{det}+ @{adj}? @{N}	[-hubieran][+hubiera]
hubiesen @{det}+ @{adj}? @{N}	[-hubiesen][+hubiese]
hubieron @{det}+ @{adj}? @{N}	[-hubieron][+hubo]
hubieren @{det}+ @{adj}? @{N}	[-hubieren][+hubiere]

Junto con esta versión de proyección extendida del grupo nominal, se han codificado otros patrones que ofrecen la opción de que aparezca un pronombre;

habían @{det}+ @{adj}? @{pn}	**habían todos los vuestros.*

Con este último patrón se pretende dar cobertura a casos similares a **Hubieron pocos muchachos en las fiestas: *los hubieron.*

Como se observará, en algunas baterías de patrones ha sido necesaria la inclusión de restricciones —exclusión del participio—, para evitar falsos positivos como *Lo habían dicho; Habéis visto la película; Los pecadores podían haber rezado.*

@{pn de CD} habrán ^({participio})	[-habrán][+habrá]
@{pn de CD} hayan ({participio})	[-hayan][+haya]
@{pn de CD} habían ^({participio})	[-habían][+había]
@{pn de CD} habrían ^({participio})	[-habrían][+habría]
@{pn de CD} hubieran ^({participio})	[-hubieran][+hubiera]

@{pn de CD} hubiesen ^({participio})	[-hubiesen][+hubiese]
@{pn de CD} hubieron ^({participio})	[-hubieron][+hubo]
@{pn de CD} hubieren ^({participio})	[-hubieren][+hubiere]

Otro de los errores frecuentes que involucran al verbo *haber* en su uso con valores existenciales es el ejemplificado por casos como **Habíamos muchos fans en el concierto*. Siendo impersonal en este uso, solo puede conjugarse en tercera persona del singular y, por lo tanto, para expresar la presencia de primeras o segundas personas, el verbo que debe utilizarse es *ser*.

habemos ^{participio}	[-habemos][+somos]
habéis ^{participio}	[-habéis][+sois]
habremos ^{participio}	[-habremos][+seremos]
habréis ^{participio}	[-habréis][+seréis]
habíamos ^{participio}	[-habíamos][+éramos]
habíais ^{participio}	[-habíais][+erais]
habríamos ^{participio}	[-habríamos][+seríamos]
habríais ^{participio}	[-habríais][+seríais]
hubiéramos ^{participio}	[-hubiéramos][+fuéramos]
hubierais ^{participio}	[-hubierais][+fuerais]
hubiésemos ^{participio}	[-hubiésemos][+fuésemos]
hubieseis ^{participio}	[-hubieseis][+fueseis]
hubimos ^{participio}	[-hubimos][+fuimos]
hubisteis ^{participio}	[-hubisteis][+fuisteis]
hubiéremos ^{participio}	[-hubiéremos][+fuéremos]
hubiereis ^{participio}	[-hubiereis][+fuereis]

Los patrones siguientes recogen la casuística generada en las perífrasis verbales en las que participa el verbo *haber; *Este año van a haber unas fiestas fantásticas; *En la final del concurso podían haber unas mil personas;*

van a haber ^{participio}	[-van][+va]
iban a haber ^{participio}	[-iban][+iba]
irán a haber ^{participio}	[-irán][+irá]

irían a haber ^{participio}	[-irían][+iría]
vayan a haber ^{participio}	[-vayan][+vaya]
fueran a haber ^{participio}	[-fueran][+fuera]
fuesen a haber ^{participio}	[-fuesen][+fuese]
fueron a haber ^{participio}	[-fueron][+fue]
fueren a haber ^{participio}	[-fueren][+fuere]

pueden haber ^{participio}	[-pueden][+puede]
podían haber ^{participio}	[-podían][+podía]
podrán haber ^{participio}	[-podrán][+podrá]
podrían haber ^{participio}	[-podrían][+podría]
puedan haber ^{participio}	[-puedan][+pueda]
pudieran haber ^{participio}	[-pudieran][+pudiera]
pudiesen haber ^{participio}	[-pudiesen][+pudiese]
pudieron haber ^{participio}	[-pudieron][+pudo]
pudieren haber ^{participio}	[-pudieren][+pudiere]

Por último, se han codificado otros cuatro patrones que abordarán oraciones erróneas en las que *hay* impersonal es acompañado, a pesar de la incompatibilidad, de pronombres reflexivos de 1ª y 2ª persona, cuando la estructura correcta exige que el verbo que complementa a *hay* se presente en 3ª persona. La presencia de estos patrones garantizará el tratamiento de oraciones como **Si queremos llegar lejos hay que esforzarnos mucho más,* que será corregida por la opción terciopersonal correcta *Hay que esforzarse mucho más.* Para tratar estos casos se han previsto los siguientes patrones:

hay @{adv}? que #{vb}@@{pn reflex}@@{1ª sg}	[-#{vb}@@{pn refl}@@{1ª sg}] [+#{vb}@@{pn refl}@@{3ª sg}]
hay @{adv}? que #{vb}@@{pn reflex}@@{1ª pl}	[-#{vb}@@{pn refl}@@{1ª pl}] [+#{vb}@@{pn refl}@@{3ª sg}]
hay @{adv}? que #{vb}@@{pn reflex}@@{2ª sg}	[-#{vb}@@{pn refl}@@{2ª sg}] [+#{vb}@@{pn refl}@@{3ª sg}]
hay @{adv}? que #{vb}@@{pn reflex}@@{2ª pl}	[-#{vb}@@{pn refl}@@{2ª pl}] [+#{vb}@@{pn refl}@@{3ª sg}]

Como se observa, la corrección prevista se basará en la supresión de los enclíticos de primera y segunda persona y la inclusión de *se*; *Hay que esforzarse mucho más.*

Otros casos de impersonalidad

Con referencia a los errores que pueden albergar otros verbos impersonales, se han etiquetado algunos patrones que corrigen expresiones erróneas generadas por la flexión de plural de estos verbos en un intento de concordancia entre el verbo y el sustantivo del sintagma preposicional; **Se trataban de muchas cosas en aquellas charlas de mujeres.*

`se #tratar##{3ª pl} @{adv}? de`	`[-#{tratar}##{3ª pl}]` `[+#{tratar}##{3ª sg}]`

Hay, además, una recomendación que se registra en PatErr con el fin de sugerir las opciones ajustadas a la norma en casos como ⊗*De aquí a Madrid se tardan dos horas.*

`se #tardar##{3ª pl} @{det}? @{N}`	`[-@{tardar}##{3ª pl}]` `[+@{tardar}##{3ª sg}]`

El programa ofrecerá como opción correctiva la propuesta por la Academia, esto es, la versión derivada de un análisis de la oración como impersonal, que exigirá que el verbo esté flexionado en tercera persona del singular.

Asimismo, se prevé tratamiento para aquellos casos en los que *bastar* se construye con la preposición *con* dando lugar a una oración impersonal que requiere que la concordancia se establezca en 3ª persona de singular; **Bastan con tres días para rematar el proyecto.*

`#bastar##{3ª pl} con`	`[-#bastar##{3ª pl}]` `[+#bastar##{3ª sg}]`

Se da cobertura, además, a aquellos casos en los que la expresión *ser suficiente* se construye con la preposición *con,* y como en el caso precedente la oración impersonal rige una concordancia en 3ª persona de singular; **Son suficientes con dos cucharadas*.

#ser##{3ª pl} @suficiente con	[-#ser##{3ª pl}] [+#ser##{3ª sg}]

13.2.3.2 *Vacilación de tiempos verbales*

Con referencia a este fenómeno de titubeo y oscilación entre las formas o modos verbales, se ha registrado un patrón que trata de subsanar casos propios del registro vulgar como **Cuando nos casemos, fuimos de viaje a París; * Ayer cantemos hasta el amanecer.* El patrón que se ha codificado tiene la forma siguiente:

{Lista 36} #{vb}##{subj}##{pres}##{1º pl}

A partir de este, y en virtud del sistema de listas planteado, el patrón tendrá una capacidad generativa que le permite ofrecer cobertura a los 13 000 verbos presentes en el lexicón. Por otro lado, la lista 36 contiene expresiones de pasado que pueden anteceder a un verbo, *ayer, antaño, hace un año, previamente*, *la semana pasada,* etc., que han sido compiladas manualmente a partir de un estudio del corpus y de las fuentes bibliográficas. La coalición de ambos recursos permite tratar aquellos casos erróneos en los que se utiliza cualquier forma verbal en primera persona de plural de presente de subjuntivo con valor de pasado, valor que será acotado por alguna de las expresiones contenidas en la lista.

La medida correctiva prevista es la sustitución de la forma verbal por la correspondiente en pretérito perfecto de indicativo.

13.3 Dequeísmo

El dequeísmo es, según los tratados tradicionales un *vicio de dicción*, en el que se hace un uso indebido de la preposición *de* que se antepone a la conjunción *que* cuando la preposición no viene exigida por ninguna palabra del enunciado.

Aunque los contextos en los que puede surgir este fenómeno son diversos, comparten la condición de que la preposición *de* no está justificada desde un punto de vista gramatical. Su inclusión errónea puede deberse bien a que se emplee la preposición cuando, por la función sintáctica de la oración, no se requiere ninguna, bien a que se utilice *de* en lugar de la preposición realmente exigida.

Para acometer toda la casuística relacionada con este fenómeno se han registrado 67 patrones codificados que se adscriben a dos incidencias escindidas; una etiqueta recoge los casos de dequeísmo en el régimen verbal general, mientras que la otra asume los errores que surgen en ese mismo entorno, pero dentro de una expresión o locución.

Se expone, a continuación, una panorámica de los contextos susceptibles de contener dequeísmo que han sido extraídos a partir del escrutinio de la teoría gramatical relacionada[96] y de los resultados que ofrece el Corpus TIP.

13.3.1 Oraciones subordinadas sustantivas

Dos de los complementos que pueden actualizarse mediante una oración subordinada sustantiva son el sujeto y el complemento directo. Ninguno de estos, que estarán encabezados por la conjunción *que,* admite la anteposición de la preposición *de.* Son incorrectas, aunque muy frecuentes, oraciones subordinadas en función de sujeto como **Me alegra de que seáis felices; *Es seguro de que nos quiere mucho; *Le preocupa de que aún no hayas llegado,* como lo son las siguientes oraciones completivas de complemento directo; **Pienso de que el final está cerca; *He oído de que te casas con Ana María.*

Del estudio de la casuística de estos contextos en los que surgen los errores, pueden extraerse un par de conclusiones relevantes para este trabajo.

[96] Náñez Fernández (1984); Almeida (2007).

En relación con las subordinadas sustantivas de sujeto, algunos de los verbos que con más frecuencia dan lugar a error son los verbos de afección, que suelen construirse con el sujeto pospuesto; *divertir, repeler,* etc.

Por otro lado, las subordinadas en función de complemento directo suelen desarrollarse con verbos de pensamiento; *pensar, opinar, creer,* etc.; de habla; *decir, comunicar, exponer,* etc.; de temor; *temer*, *maliciarse*, etc., y de percepción; *ver, advertir,* etc.

A partir de estos criterios semánticos, se ha llevado a cabo una búsqueda de verbos que presenten estos rasgos semánticos*, afección, pensamiento, habla* y *percepción* y se ha contrastado la posibilidad de que puedan participar en estructuras dequeístas. La siguiente lista ofrece los verbos susceptibles de generar este error.

adorar	creer	fascinar	opinar
amar	confesar	gustar	pensar
apetecer	decir	imaginar	repeler
asegurar	detestar	insistir	sentir
comunicar	divertir	interesar	sospechar
confiar	encantar	necesitar	temer
considerar	enternecer	odiar	ver
constar	exponer	oír	

El patrón tipo que se ha codificado es como el que sigue:

#necesitar @{adv}+? de que	[- de que][+ que]

Como se observa, se da cabida a la posibilidad de que se intercalen uno o más adverbios entre el verbo y la secuencia *de que.* De este modo se flexibiliza el patrón para aumentar la cobertura; **Necesito urgentemente de que vengas; *Detesto absolutamente de que me mientas.*

Tras comprobar la viabilidad de estos patrones y su funcionamiento en textos reales se ha hecho necesario, en muchos casos, marcar limites que deben imponerse a la aplicabilidad generalizada de algunos patrones. En el campo de la

base de datos consignado a las condiciones del patrón, se hacen estas observaciones que deben asumirse y ejecutarse previa a la aplicación del patrón.

Consideraciones previas a la aplicación del patrón

En algunos casos, como se ha observado en la generación del patrón del verbo *decir,* es necesario hacer ciertas restricciones para evitar que la ambigüedad que se genera con las formas verbales derivadas, —en este caso con el participio *dicho*—, pueda dar lugar a falsos positivos como *El dicho de que no hay que fiarse de nadie es muy cierto.*

Es ineludible, por lo tanto, realizar un estudio de cada verbo codificado con el fin de localizar las formas ambiguas que puedan desarrollar otras categorías gramaticales. De este modo se ha comprobado la validez de cada patrón y, en los casos necesarios, se han excluido las formas ambiguas que comprometían el correcto tratamiento del error.

Otro patrón que ha necesitado acotaciones es el del verbo *asegurar/asegurarse*.

`#asegurar @{adv}+? de que`	`[- de que][+ que]`

Es necesario excluir de este patrón los casos en que este verbo presente pronombres clíticos, para evitar correcciones inadecuadas en oraciones como *Asegúrate de que venga; Hay que asegurarse de que venga; Aseguraos de que venga.* Por otro lado, es necesario vetar la aplicación del patrón cuando este verbo se encuentre precedido de la expresión *lo que* más un posible pronombre átono o la estructura paralela «demostr + *que* + pn átono?». De este modo se evita la corrección indebida de *Lo que te aseguré de que vendría era una suposición.*

Otros verbos como *oír, comunicar, imaginar, temer, sospechar,* etc. presentan también estos mismos límites; *Aquello que (te) imaginabas de que nos íbamos de vacaciones; Lo que (me) comunicaron de que nos darían el visto bueno*.

Por último, debe hacerse una consideración relevante con respecto a algunos verbos de afección que muestran alternancia entre complementos, sujeto o complemento de régimen. Estos verbos pueden generar dequeísmo cuando se utiliza la tercera persona de cualquier tiempo y modo seguido de *de que*, haciendo funcionar esta subordinada sustantiva —siempre en modo subjuntivo— como complemento de régimen en lugar como de sujeto. Estos verbos propician pares como los siguientes:

Me alegro de que vengáis.
**Me alegra de que vengáis;* dequeísmo
Me alegra que vengáis.
**Me alegro que vengáis;* queísmo

Como se explica en la NGLE, estas construcciones surgen como un intento por parte del hablante de dejar una marca formal en el discurso *para introducir los complementos y sujetos oracionales posverbales ante determinados predicados* (§43.6d).

Los verbos pronominales que dan lugar a estos casos acotados —en tercera persona— de dequeísmo son los siguientes:

preocuparse	avergonzarse	entristecerse
alegrarse	ofenderse	dolerse
avergonzarse	convencerse	congratularse
lamentarse	aburrirse	apenarse
cansarse		

Los patrones que se han codificado para estos casos son de dos tipos. El primero de ellos corregirá el dequeísmo, el segundo subsanará el queísmo en los casos en los que se omita la preposición *de* cuando sea necesaria.

`@{pn}@@{reflex}^se #sorprender##{3ª} @{adv}+? de que`	`[- de que]` `[+ que]`

El patrón se encabeza con la restricción en el pronombre que precede la secuencia para evitar falsos positivos como *Aún se sorprende mucho de que su suerte no le acompañe.*

Por otro lado, como en casos anteriores, este patrón tiene ciertas condiciones de aplicabilidad en cuanto al cotexto que le precede; este nunca podrá estar representado por secuencias como «*lo que* + pn átono opcional» o «demost + *que* + pn átono opcional», para eludir correcciones desacertadas en oraciones correctas como: *Lo que me sorprende de que luche es que no va a conseguir nada; Aquello que le avergonzaba de que siempre estuviera sola…*

`#sorprenderse^##{3ª}) @{adv}? que`	`[- que][+ de que]`

Con este patrón se trata un fenómeno anejo, el queísmo, que se desarrollará en el próximo epígrafe. Se corrigen a partir de estos patrones, casos como **Me sorprendo que aún no haya dado noticias de vida; *Es curioso que os sorprendáis exclusivamente que no os devuelvan vuestros gastos*.

13.3.1.1 *Algunas expresiones copulativas*

El dequeísmo asociado a las oraciones subordinadas sustantivas encuentra otro contexto donde poder desarrollarse; **Es posible de que pierda los nervios de un momento a otro.* Con frecuencia se registran casos de dequeísmo en estructuras copulativas que tienen como sujeto una subordinada sustantiva prologada por *que.*

Los resultados de la investigación a partir de los datos ofrecidos por el Corpus TIP se actualizan en una lista que contiene las expresiones más frecuentes con la estructura «*ser +*adj» que toman una subordinada de sustantivo como sujeto:

```
ser cierto de que
ser posible de que
ser probable de que
ser interesante de que
ser fácil de que
ser difícil de que
ser necesario de que
ser urgente de que
ser seguro de que
```

Como en los casos anteriores, en el patrón se prevé la posibilidad de que en la secuencia se inserte uno o más adverbios; **Es mucho más fácil de que suban los sueldos a (de) que bajen los precios; *Es muy necesario de que confíes en mí.*

#ser fácil @{adv}? de que	[- de que][+ que]

13.3.1.2 *Oraciones subordinadas sustantivas con función de atributo*

Otro contexto en el que se pueden constatar casos de dequeísmo es en los márgenes del atributo. Cuando se antepone la preposición *de* a una oración subordinada que ejerce las funciones de este complemento en oraciones copulativas con el verbo *ser* es común que surjan estos errores. Este complemento, por lo general, no va precedido de preposición alguna y, por lo tanto, serán incorrectas oraciones como: **Mi intención es de que participemos todos en la rifa; No parece de que todos estén de acuerdo con las medidas del ministro.* Se codifican, asociados a este contexto, los siguientes patrones:

#parecer de que	#ser de que	[- de que][+ que]

13.3.1.3 *Verbos con vacilación en su rección preposicional*

Otra de las fuentes de error que puede derivar en dequeísmo se encuentra en los verbos que, por vacilación entre preposiciones —error en su régimen preposicional— se acompañan de la preposición *de* indebidamente. Así se observan casos como **Insistieron de que fuéramos con ellos; *Me fijé de que llevaba corbata*.

Teniendo en cuenta estos casos, se han incluido en PatErr los verbos que exigen régimen preposicional y resultan problemáticos en relación con este fenómeno.

insistir de que	quedar de que
confiar de que	estar de acuerdo de que
fijarse de que	coincidir de que

Los patrones codificados seguirán el siguiente modelo:

#insistir @{adv}? de que	[- en que][+ que]

13.3.2 Locuciones conjuntivas

Se ha observado, por último, la inserción indebida de la preposición *de* en algunas locuciones conjuntivas que no la exigen, posiblemente por similitud con otras locuciones que sí la precisan —*antes de que, después de que, a condición de que, con tal de que,* etc.—. En relación con este contexto, se han codificado patrones para las siguientes expresiones:

```
a no ser de que                    al punto de que
a medida de que                    a menos de que
desde luego de que                 dado de que
una vez de que
```

a medida de que	[- de que][+ que]

13.4Queísmo

El desarrollo de este fenómeno concluye en la supresión indebida de una preposición —generalmente *de*— ante la conjunción *que,* cuando la preposición es requerida por alguna palabra del enunciado.

Este fenómeno es cada vez más frecuente como manifestación de una ultracorrección debido al temor a cometer dequeísmo, por lo que se suprime la preposición cuando el nexo *de que* es efectivamente preceptivo.

Tras un examen de los contextos y cotextos en los que se constatan los casos de queísmo más frecuente, se observa la omisión de la preposición en ciertos verbos que necesitan, para completarse, un complemento de régimen. Es el caso de *acordarse*, *alegrarse*, *arrepentirse*, *olvidarse*… *de*. Es común, también, la supresión de la preposición en sintagmas cuyos núcleos —sustantivos o adjetivos— piden un complemento preposicional; *a condición de*, *con ganas de, seguro de*, *convencido de*, etc. Un último contexto en el que se puede verificar la presencia de este error es en algunas locuciones conjuntivas como *a pesar de*, *a fin de*, *a condición de*, *en caso de*, etc.

Se ha llevado a cabo un estudio sobre los límites y agentes de este fenómeno que ha concluido en el registro de 49 patrones organizados en baterías según sea la preposición suprimida que genera el queísmo.

13.4.1 Supresión de la preposición *de*

13.4.1.1 *Queísmo en régimen verbal general*

Se recogen los verbos que se ha constatado que con frecuencia generan este error.

acordarse	cuidar	lamentarse
tratar	olvidarse	arrepentirse

El patrón que se ha codificado para estos casos será del tipo:

#lamentarse @{adv}? que	[- que][+ de que]

De nuevo, debe tenerse especial atención con relación a los casos de ambigüedad; así en el patrón dedicado al verbo *tratar* o *cuidar* debe ponerse como condición que se excluyan las formas verbales ambiguas para evitar la corrección de secuencias correctas como; *El trato que me dio no es el que me esperaba; El cuidado que me prometió fue efímero.* Estas consideraciones se encuentran explicitadas en el campo de la base de datos consignado a la condición de aplicabilidad del patrón.

Dentro de este bloque, hay un subgrupo que puede subsumirse bajo la premisa de ser verbos pronominales que se construyen con un complemento de régimen. Tras su estudio, se observa la peculiaridad de que deben excluirse de este grupo las terceras personas de todos los verbos para evitar la corrección de secuencias como: *Me preocupa que nada vuelva a ser como antes; Me alegró que llegaras sin avisar.*

preocuparse	avergonzarse	cansarse
ofenderse	congratularse	entristecerse
alegrarse	apenarse	aburrirse
sorprenderse	convencerse	

El patrón modelo será:

#entristecerse^##{3ª} @{adv}? que	[- que][+ de que]

y solventará casos erróneos como **Me entristezco que no llegue el paquete.*

13.4.1.2 *Nexos subordinantes y locuciones conjuntivas*

Las partículas que tras un estudio de contextos se han revelado como susceptibles de ser afectadas por este fenómeno son las siguientes:

a pesar	a expensas	en vista
a fin	a propósito	hasta el punto
a cambio	con la condición	por razón
a condición	en caso	
a causa	en virtud	

además	encima
aparte	enseguida

#en virtud que	[- que][+ de que]

13.4.1.3 *Expresiones o perífrasis verbales*

darse la casualidad	caer en la cuenta
dar la impresión	no haber duda
darse cuenta	no caber duda
arder en deseos	tener ganas

El patrón flexible previsto para estos casos será similar al siguiente:

#dar @{det}? @{adj}? @casualidad que	[- que][+ de que]

Debe tenerse en cuenta en casos como el expuesto o en el que se involucre el sustantivo *duda*, que estos sustantivos pueden encontrarse en plural; **No caben dudas que llegará con retraso.*

Por otro lado, se ha contrastado que algunas de estas expresiones tienen que acotarse con un verbo en subjuntivo tras el *que*,

```
#arder @{adv}+? en deseos que #{vb}##{subjuntivo}
```

para evitar falsos positivos como; *Ardo en deseos de verte* y corregirlo por **Ardo en deseos de que verte.*

13.4.1.4 *Expresiones con sustantivos*

Es común escuchar o leer expresiones como **Te dejo ir con la condición que no me olvides; *Le aumentaron el sueldo con la condición que trabaje más horas.* Para solventar estos casos, se han codificado dos patrones con las siguientes expresiones:

```
con la condición        a condición
```

Los patrones que codifican estas expresiones también deben incluir el uso de un subjuntivo tras la conjunción para evitar la corrección de secuencias aceptables como; *Con la condición que te han impuesto es complicado que vuelvas a la empresa.*

`con la @{adj}? condición que #{vb}##{subjuntivo}`	`[- que]` `[+ de que]`
`a condición que #{vb}##{subjuntivo}`	`[- que]` `[+ de que]`

13.4.1.5 *Expresiones con adjetivos*

Como en el caso precedente, algunas estructuras copulativas en las que participan ciertos adjetivos pueden dar lugar a errores de queísmo.

```
estar convencido     estar seguro       ser consciente
```

La codificación de estos patrones es del tipo:

`#ser @{adv}+? @consciente que`	`[- que][+ de que]`

Nótese que se posibilita la flexión de los adjetivos para dar cabida a casos en los que se presente el error con esta forma flexionada; **Estamos absolutamente seguras que es una buena compra.*

13.4.2 Supresión indebida de otras preposiciones

El estudio de la casuística derivada por estos fenómenos ha revelado otras incidencias provocadas por la supresión de una preposición diferente a *de.* Se observa este fenómeno en algunos verbos no pronominales que se construyen con complemento de régimen cuyo enlace es la preposición *en*.

```
confiar (en) que                insistir (en) que
fijar (en) que                  estar de acuerdo (en) que
```

El patrón previsto para estos casos será del tipo:

`#insistir @{adv}+? que`	`[- que][+ en que]`

13.5 Régimen preposicional

Este bloque recoge algunos casos en los que se ha constatado que la incidencia está provocada por cuestiones relacionadas con el régimen preposicional que presentan ciertos verbos o expresiones.

13.5.1 Errores de régimen preposicional en términos o expresiones

13.5.1.1 **adicción con, *opción a, *idéntico con*

No es extraño encontrarse en un texto con expresiones incorrectas del tipo **Tengo adicción con los videojuegos; *Soy adicta de internet.* El complemento de la palabra

adicción y el de sus derivados *adicto* suele ir introducido por la preposición *a* o, menos frecuentemente por *por*.

En estos casos, y en todos sus derivados, se extraerá la preposición inadecuada *con* y se sustituirá, en primera instancia, por *a*, aunque el programa también tiene prevista la opción de sustitución por la preposición *por*.

Sucede lo mismo con el adjetivo *idéntico* que con frecuencia se hace acompañar de *con* o *que* en lugar de *a,* o con los sustantivos *dolor* y *opción* en casos como **No tenía opción a ganar, *No aguanto más el dolor a la cabeza que tengo.* En todos estos casos, que pasan desapercibidos para otros correctores, el programa sustituirá la preposición inadecuada por la correcta.

13.5.1.2 **mayor a, *mayor de*

Es común la escritura de expresiones erróneas como **El tiempo de ejecución es mayor a/de lo esperado.*

Mayor, es un adjetivo comparativo sincrético que exige que el término de la comparación que le sigue vaya encabezado por *que.* El mismo régimen debe aplicarse a otros comparativos sincréticos como *menor, mejor* y *peor.*

Para solventar este fenómeno se han codificado seis registros por adjetivo, tres en los que el patrón presenta la opción incorrecta con la preposición *a* y otros tres para los casos con *de*:

`@mejor de @{N}@@{prop}`	`@mejor a @{N}@@{prop}`
**No eres mejor de/a Manuel.*	

`@mejor de @{pn}@@{sujeto}`	`@mejor a @{pn}@@{sujeto}`
**Lo hice mejor de/a ella.*	

`@mejor de @{det}@@{art}`	`@mejor a @{det}@@{art}`
**Fue mejor de/a lo que esperaba.*	

En todos ellos la acción correctora es la misma; sustitución de la preposición que sigue al adjetivo por la conjunción *que*.

13.5.1.3 **a la mayor brevedad posible*

Junto con estos términos y expresiones se ha codificado el patrón de error **a la mayor brevedad posible*, en el que se cambia la preposición que encabeza la expresión por *con*. A pesar de no constituir un error, la norma prefiere y recomienda la versión que comienza con la preposición *con*.

13.5.2 Errores de régimen preposicional en verbos

Asociados a este fenómeno, se han codificado once patrones gracias a los cuales pueden tratarse ciertos errores que surgen a propósito del régimen preposicional que exigen algunos verbos. Los verbos que con más frecuencia presentan estas incidencias quedan recogidos en los siguientes patrones:

#divergir con	[- con][+ de]|[+ en]
#dignarse en	[- en][+ a]
#hartar a	[- a][+ de]
#enfrentar contra	[-contra][+con]
#encarar contra	[-contra][+con]
{numeral} partido de {numeral}	[-partido de][+partido por]
#ganar de @{cuantificador}	[- de][+ por]
#ganar de {numeral}	[- de][+ por]
#colaborar a	[- a][+ en]| [+ con]
#enfrentar a	[- a][+ con]|[+ contra]
#encarar a	[- a][+ con]

13.5.3 Errores con verbos que rechazan preposición

En este bloque se recogen los verbos que, aunque erróneamente, con frecuencia son escritos seguidos de una preposición. En los casos estudiados, esta preposición que se cuela indebidamente es *de* y genera expresiones que pueden considerarse propias del registro vulgar. Los verbos que se han codificado en patrones son los siguientes;

```
hacer de sufrir                    hacer de rabiar
hacer de llorar                    intentar de
hacer de reír
```

`@{pn}@@{reflex} #recordar de`	`[-#recordar][+#acordar]`

13.5.4 Avisos de régimen preposicional en verbos

Un tratamiento diferenciado merece el verbo cesar, cuyo uso suele asociarse al error. Se han dedicado dos patrones para advertir al usuario del régimen preposicional que presenta este verbo.

`#cesar del @{det}+? @{adj}? @{N}`	`[- del ][+ en ]`
`#cesar de @{det}+? @{adj}? @{N}`	`[- del ][+ en el ]`

A diferencia de *dimitir*, cuando este verbo está seguido de un sustantivo y tiene el significado de "dejar de desempeñar un cargo", la preposición que debe seguirle es *en*; *Ha cesado en su cargo.*

El aviso lingüístico aparecerá, pues, ante secuencias como **Nuestros políticos no cesan de sus privilegiados cargos,* y será necesario llevar a cabo una corrección interactiva. La polisemia de este verbo impide que pueda fijarse un patrón de error en lugar de uno de aviso. De ser así, el programa corregiría siempre de modo automático la preposición *de* por *en*, y trataría casos correctos como: *Mi tío llegó a casa hablando sin cesar de la cuestión; Tenía la obligación de cesar de inmediato su mal comportamiento,* convirtiendo finalmente, oraciones correctas en errores.

Capítulo 14

Léxico

Los casos que se presentan a continuación, junto con otros tantos que están codificados en PatErr y no serán expuestos aquí, constituyen errores y estructuras erróneas generados por asuntos relacionados con el léxico o la semántica. La selección de los casos tratados obedece como en el resto de los epígrafes, a la amplia presencia de estos errores en los textos escritos del español actual. Esta frecuencia no obsta para que la gran mayoría de los casos no sean identificados ni tratados por ningún corrector de los disponibles.

Como cabe esperar, las listas de palabras o expresiones que se presentan son listas iniciales e ilustrativas de cómo llevar a cabo el tratamiento de esos fenómenos. No son, por lo tanto, listas finalizadas y, como el repertorio léxico de una lengua, deberán estar en perpetuo crecimiento. Tanto los neologismos, como los extranjerismos o las imprecisiones léxicas forman paradigmas en constante progresión y enriquecimiento.

Los fenómenos que se abordarán en este capítulo tendrán relación con la precisión de algunos términos y las imprecisiones léxicas, los neologismos y la redundancia.

14.1 Impropiedades léxicas

En este epígrafe se pretende dar una muestra del tratamiento que propone PatErr para términos cuya utilización suele constituir un error que deviene en una impropiedad léxica, entendida esta como una falta de propiedad y pericia en el uso intencionado de una palabra, generalmente como consecuencia de una atribución errónea de significado.

Son más de 30 los casos codificados para resolver cientos de errores, que en su mayoría, se concentran en el registro periodístico y en el administrativo. Citamos algunos de los casos más productivos en los que con frecuencia se registra un uso impropio.

14.1.1 *Cesar*, *dimitir* y *destituir*

Esta tripla de verbos, cuyos significados y reacción preposicional se intercalan hasta desbarrarse, participan con relativa frecuencia de expresiones que presentan impropiedad en el plano del significado e inconsistencias en el sintáctico; **Juana ha cesado en el cargo; *Del Bosque fue dimitido tras un trabajo espectacular; *El ministro dimitido no tiene, por el momento, sucesor.*

Para desenmarañar el entramado de significados y preposiciones que pueden adoptar estas formas se ha formalizado el siguiente conjunto de patrones:

#ser @dimitido[97]	[- @dimitido] [+ @destituido]	*Dimitir* significa renunciar o hacer dejación de algo, por lo tanto, es impropia la construcción **ser dimitido de algo por alguien*. En su lugar, debe utilizarse ser *destituido* o *cesado*.
@{N} @dimitido	[- @dimitido] [+ @dimisionario]	El participio *dimitido* no debe utilizarse como adjetivo para hacer referencia a la persona que dimite. En su lugar, se utiliza el término *dimisionario*: *El secretario dimisionario ha sido sustituido por una mujer muy competente.*
@dimitido @{N}	[-@dimitido] [+@dimisionario]	
#cesar a	[-@cesar] [+@destituir]	*Cesar* significa dejar de desempeñar un cargo, por lo que *cesar a alguien* es incorrecto; en su lugar se recomienda utilizar *destituir* a alguien: *Se destituyó al concejal de festejos por aburrido.*
#ser @cesado	[- @cesado] [+ @destituido]	

Se suman a este bloque dos avisos lingüísticos que aportan información sobre la naturaleza semántica y la rección preposicional del verbo *cesar:*

[97] A pesar de que *dimitido* es un participio en esta estructura y que no debería presentar capacidad flexiva, se codifica aquí como adjetivo con el fin de abarcar las formas *dimitida, dimitidísimo,* etc.

#cesar de @{det}+? @{adj}? @{N}	[- de] [+ en]	A diferencia de *dimitir*, cuando a este verbo le sigue un nombre con el sentido de dejar de desempeñar un cargo, la preposición que rige es *en*: *Ha cesado en su cargo.* Cuando se hace explícito el cargo que se deja se prefiere la forma *como*: *Ha cesado como presidente.*
#cesar del @{det}+? @{adj}? @{N}[98]	[- del] [+ en el]	

14.1.2 *Cuyo* despojado de posesividad

Es muy frecuente toparse con secuencias en las que el determinante posesivo *cuyo* —y sus variantes— se presenta desprovisto de ese rasgo semántico de pertenencia. Expresiones desarrolladas por sintagmas preposicionales como *en cuyo caso, por cuya causa, con cuyo fin*, en muchas ocasiones y contextos han sacrificado su valor posesivo en favor de uno más cercano al de los demostrativos.

Ante estos usos impropios desaconsejados por la norma, se ha optado por diseñar una recomendación que recuerde la naturaleza posesiva de esta forma y el sentido que debe presentar en estas construcciones preposicionales prácticamente lexicalizadas.

{prep} cuya causa {prep} cuyas causas	[-cuya] [+tal]|[+esa]	No es apropiado el uso de *cuyo* y sus derivados en expresiones en las que no presenta valor posesivo. Expresiones como *en cuyo caso*, *con cuyo fin*, no están recomendadas por la norma debido a que el pronombre posesivo no aporta este significado sino más bien el de demostrativo. Se prefieren las expresiones *en tal caso*, *en ese caso*, o los relativos *el que*, *el cual*, etc.

Se han codificado, además, otros sustantivos que suelen protagonizar estas secuencias:

[98] Este último patrón amplía la cobertura al dar cabida a la contracción *del* que modificará, en consecuencia, la secuencia de la corrección.

{prep} cuyo caso {prep} cuyos casos	{prep} cuyo motivo {prep} cuyos motivos
{prep} cuyo fin {prep} cuyos fines	{prep} cuyo objeto {prep} cuyos objetos

Como se observa, estos patrones no se han agrupado mediante marcas de codificación en el determinante y sustantivo para dar cabida a la secuencia en plural. Las flexiones que derivarían de los dos elementos serían prescindibles en una gran parte y generarían patrones inútiles en los que, por ejemplo, habría errores de concordancia; **cuya casitos*. Este caso ilustra, a pequeña escala, lo que se señaló en el planteamiento de PatErr con relación al problema de sobredimensionar un error.

14.1.3 Restricciones semánticas e impropiedades en algunos verbos

Son muchas las formas que consignan restricciones sintácticas, temáticas o semánticas que no se contemplan en el discurso. Se ha codificado, a este propósito, un grupo de verbos que suelen emplearse impropiamente. Para su tratamiento, y ante las limitaciones semánticas que se imponen a las técnicas que desarrollan este recurso, se han diseñado avisos lingüísticos que informan sobre el régimen inherente a estas formas.

Proponemos, como modelo, el verbo *adolecer:*

#adolecer	Este verbo significa *tener o padecer algún defecto, dolencia o enfermedad.* Exige estar seguido por un complemento encabezado por *de,* que hace referencia a la dolencia en cuestión. Es, por tanto, impropio utilizar este verbo con el significado de *carecer*, *estar falto,* cuando esa ausencia hace referencia a algo positivo: ⊗*Adolecen del cariño que se merecen.* En este sentido, es preferible utilizar la expresión *estar falto de* o *carecer*.

Este mismo patrón y tratamiento se ha previsto para el siguiente grupo de verbos *conflictivos*:

```
#atesorar        @{pn}@@{reflex} #dignar   #involucrar    #ostentar
#barajar         #hacer gala              #infligir      #tachar de
#detentar        #interceptar             #infringir     #tildar de
```

Otros de los varios casos registrados que serán objeto de aviso lingüístico serán pares de adjetivos como *eficaz* y *eficiente*, el impropio uso frecuente de *bizarro* por *extravagante,* el de *deleznable* por *detestable* o el de expresiones mal empleadas como *solución de continuidad.*

14.2 Precisión

Dentro del tratamiento que PatErr puede ofrecer para temas relacionados con las impropiedades léxicas, cabe hacer mención a un grupo de términos y expresiones para los que se han diseñado avisos lingüísticos. En estas glosas se precisa el significado de estas formas que, con frecuencia, son utilizadas al margen de ciertos matices semánticos relevantes.

Se han codificado los siguientes pares de adjetivos que parecen ser sinónimos, pero no lo son y suelen aparecer en los textos indistintamente.

@israelí @israelita	El adjetivo *israelita* es sinónimo de *hebreo o judío*, tanto en sentido histórico como religioso. Para hacer referencia al topónimo del moderno estado de Israel, el adjetivo apropiado es *israelí*. De este modo, hablaremos del *actual presidente israelí*, y del *arte israelita*, como sinónimo de arte hebreo o judío.
@islámico @islamista	El adjetivo *islamista* se asocia a algo perteneciente o relativo al integrismo musulmán en la esfera política o social. Para una definición más genérica, del islam o que profesa el islam, el término adecuado es *islámico*. De este modo, nos referiremos a un *régimen islamista*, pero no a una *música islamista* sino *islámica*.
@termal @térmico	Este adjetivo deriva del sustantivo *termas*, baños públicos de los antiguos romanos. Para hacer referencia a algo perteneciente o relativo al calor o la temperatura, el adjetivo correcto es *térmico*. Bajo esta pauta, hablaremos de *baños termales* y de *centrales térmicas*.

Con el mismo criterio se codifica la siguiente batería;

Iberoamérica Latinoamérica Hispanoamérica Sudamérica Suramérica	*Suramérica*, sinónimo de *Sudamérica*, hace referencia al entorno geográfico y engloba los países situados al sur del istmo de Panamá. Otros términos relacionados son: *Hispanoamérica* conjunto de países americanos que tienen lengua española; *Latinoamérica*, en referencia al conjunto de países que tienen lenguas derivadas del latín (español, portugués y francés), e *Iberoamérica* que alude al conjunto de países que tienen habla española o portuguesa.

Se aborda, por último, la distancia semántica que media en pares de expresiones como *deber* y *deber de*.

#deber de {inf}	Esta construcción verbal se utiliza para expresar probabilidad, incertidumbre o duda. Cuando el contexto resulta suficientemente claro, se puede suprimir la preposición *de*. Para la expresión de obligación con este verbo, es necesario eliminar la preposición *de*. ⊗*Suso debió de ir a ver a su madre en aquel momento dramático.* *La sandía debe de pesar unos 10 kg.*
#deber {inf}	La perífrasis compuesta por *deber* y un infinitivo se utiliza para expresar obligación. Cuando el sentido que se quiere expresar es de posibilidad o duda es necesario acompañarlo de la preposición *de*. ⊗*Debería haber ganado mucho dinero porque se ha comprado un cochazo.* *Todos deberíamos concienciarnos con su causa.*

14.3 Neologismos

Se ha desarrollado un pequeño grupo de patrones de aviso lingüístico que aportan información semántica o pragmática relacionada con la adecuación y pertinencia en ciertos registros de estos términos nuevos en el discurso del español.

Se han recogido términos neológicos, adjetivos y sustantivos en su mayoría, que resultan imprecisos, incoherentes o innecesarios en nuestra lengua. Algunos como *antiedad* o *rejuvenecedor* son propios del ámbito publicitario y/o del bienestar, otros como *monitoreo* o *posicionamiento* se despliegan en el argot de las

nuevas tecnologías. Junto a estos términos, se han codificado algunos neologismos verbales especialmente frecuentes en el español de América[99].

#inicializar	#iniciar	Este verbo, propio del contexto de la informática, significa *establecer los valores iniciales para la ejecución de un programa*. Fuera de este contexto, su uso no está recomendado y se prefiere el verbo *iniciar*.
#monitorear	#monitorizar	En la variedad de español de España, se prefiere la forma *monitorizar* para hacer referencia a la *observar mediante aparatos especiales el curso de uno o varios parámetros fisiológicos o de otra naturaleza.*
#recepcionar	#recibir	En la variedad de español de España, no es adecuado utilizar esta forma en el sentido de *recibir*. En el español de gran parte de América adquiere unos valores precisos que limitan el uso de este verbo.

14.4 Redundancia

Entendemos aquí la redundancia como una propiedad que presentan algunos mensajes en los que se desarrolla una repetición innecesaria de información para expresar una idea o concepto que ha quedado ya manifiesto en el contenido del mensaje.

Estos errores pueden ser analizados o clasificados bien desde una perspectiva estilística, bien desde un enfoque léxico-semántico. Mientras que expresiones como ⊗*Llegaron a la misma vez,* parecen afectar más a cuestiones de naturaleza estilística, ejemplos como ⊗*Subió para arriba en cuanto lo supo*, pueden ser considerados como imprecisiones o equívocos desarrollados en el plano léxico-semántico. En cualquier caso, y como puede advertirse, los límites de este fenómeno se presentan difusos y escurridizos cuando lo que se pretende es anclar el fenómeno a uno de los niveles de la lengua.

Para dar cobertura a estos errores de naturaleza léxica y estilística se han codificado 40 patrones de error. Estos pueden dividirse en dos bloques; uno que se

[99] Estos datos de carácter diatópico están anotados en el campo correspondiente a VARIEDAD en la base de datos, y constituyen rasgos que pueden resultar útiles para la definición y desarrollo de perfiles de usuario.

diseña para tratar expresiones y estructuras genéricas, y otro que aglomera expresiones idiomáticas frecuentes que presentan signos de redundancia.

14.4.1 Estructuras genéricas

Se ha desarrollado un manojo de patrones que detectan estructuras en las que típicamente se observan expresiones redundantes, solo aceptables cuando la redundancia esté deliberadamente buscada por el autor con fines intensificadores o humorísticos.

Al margen de usos y giros estilísticos, es frecuente encontrar, tanto en la lengua hablada como en la escritura, la combinación de adverbios cuantificadores o de grado —*muy, poco, bastante, demasiado,* etc.— con las formas superlativas del adjetivo; ⊗*demasiado orgullosísimo,* ⊗*bastante cansadísimo,* ⊗*poco elegantísima,* ⊗*muy enorme,* ⊗*extremadamente gélida,* etc.

Para evitar estas coapariciones en las que por la naturaleza semántica de los términos surge la redundancia, se han desarrollado los siguientes patrones:

`{adv de grado} @{adj}@@{superlativo}`	`[-{adv de grado} ]`
`{adv de grado} @{adj}@@{superlativo}`	`[-@{adj}@@{superlativo}]` `[+@{adj}]`

Como se observa, el tratamiento que ofrecen estos patrones para evitar la redundancia se basa en dos recursos; la supresión del adverbio cuantificador o la supresión del sufijo de superlativo adherido al adjetivo, ambos con aportaciones semánticas —de grado sumo— similares.

Otro contexto en el que se registran redundancias es aquel en el que un adverbio de grado modifica a un adjetivo cuya cualidad queda expresada en su grado más extremo sin necesidad de flexión superlativa; ⊗*muy atroz,* ⊗*casi abominable,* ⊗*poco terrible,* etc.

`@{adv}@@{grado} @{adj}@@{grado extremo}`	`[-@{adv de grado} ]`

Para la codificación de estas baterías de patrones genéricos ha sido necesaria, tras un estudio de corte semántico, la creación de dos repertorios etiquetados que recogen, por un lado, los adverbios de grado, y por otro, los adjetivos de grado extremo.

Para ilustrar el contenido de estas listas compiladas manualmente en Paterr se ofrecen los anexos 13 y 14. En la primera, queda establecido el elenco de adverbios de grado del tipo *sumamente, absolutamente, extremadamente, muy, bastante, tan, tanto,* etc. El repertorio, en este punto de desarrollo, cuenta con 35 términos. Por otro lado, el Anexo 14 que contiene la Lista 7, ofrece un paradigma con 44 adjetivos de grado extremo que por su naturaleza semántica vetan la modificación de un adverbio de grado. Ejemplos de este paradigma son los adjetivos *brutal, colosal, delicioso, encantador, enorme, esencial, espantoso*, etc.

Los elementos que constituyen estas listas, que se etiquetarán en el Lexicón TIP con estos rasgos semánticos, serán reutilizados agrupados así para la configuración de otros patrones.

Otras expresiones de uso frecuente que presentan redundancia son aquellas similares a ⊗*Mi perro es más mayor que el tuyo*; ⊗*Te dediqué los más mejores años de mi vida*; ⊗*El más peor momento de mi vida lo pasé contigo*.

Para evitar la concurrencia del adverbio *más* y las formas sincréticas del comparativo de superioridad de algunos adjetivos, que presentan inherentemente en su significado ese valor del *más*, se han diseñado los siguientes patrones:

`más @mayor que`	`[-más ]`
`más @mejor que`	`[-más ]`
`más @menor que`	`[-más ]`
`más @peor que`	`[-más ]`

Tanto *mayor*, como *mejor, menor* y *peor*, que serán codificados en patrones independientes replicando el modelo expuesto, son adjetivos sincréticos de grado

superior de *grande, bueno, pequeño* y *malo*, respectivamente y rechazan la modificación o graduación del adverbio *más* por generar expresiones redundantes.

14.4.2 Expresiones redundantes frecuentes

Se ofrece, a continuación, un conjunto de expresiones que se constatan con frecuencia y manifiestan errores de redundancia. En la mayoría de estos casos, la acción correctora se basa en la supresión de la segunda parte de la expresión, que resulta prescindible en términos de contenido; ⊗*lendakari vasco,* ⊗*nexo de unión,* ⊗*subir para arriba,* ⊗*deambular sin rumbo, etc.*

@accidente @fortuito	[- @fortuito]
@antecedente @previo	[- @previo]
@aterido de frío	[- de frío]
@autopsia a los cadáveres	[- a los cadáveres]
@autopsia al cadáver	[- al cadáver]
@bajar para abajo	[- para abajo]
@cáncer @maligno	[- @maligno]
@cita @previa	[- @previa]
@colofón @final	[- @final]
@conllevar consigo	[- consigo]
@constelación de estrellas	[- de estrellas]
@copar por completo	[- por completo]
@deambular sin rumbo	[- sin rumbo]
@entrar adentro	[- adentro]
@erradicar de raíz	[- de raíz]
@falso @pretexto	[-@falso]
@funcionario @público	[- @público]
@homosexual y @lesbiana	[- y @lesbiana]
@interrelacionar entre sí	[- entre sí]
@lapso de tiempo	[- de tiempo]
@lendakari @vasco	[- @vasco]
@nexo de unión	[- de unión]
@opción @alternativa	[- @alternativa]
@precedente @previo	[- @previo]
@prever con antelación	[- con antelación]

@salir para? afuera	[- para? afuera]
@subir para arriba	[- para arriba]
a la misma vez	a la vez
actualmente en vigor	en vigor
mayor brevedad posible	mayor brevedad
mayor celeridad posible	mayor celeridad

Por último, se abordan un par de expresiones frecuentes que revelan redundancia informativa debido a que los dos términos que constituyen la expresión presentan el mismo significado:

ambas dos	las dos
ambos dos	los dos

Capítulo 15

Estilo

En este capítulo se ofrece una muestra ilustrativa del tratamiento y contenido al que, en materia de estilo, ofrece cobertura el recurso PatErr. Como sucede con otros fenómenos, una gran parte de los patrones que siguen esta lógica e intención —corrección estilística— estarán diseminados en otros epígrafes como el asignado a la construcción de una expresión, que se expondrá en el próximo capítulo, a la utilización de los neologismos, vulgarismos, extranjerismos o aquellos que tratan asuntos gramaticales como la concordancia y la rección preposicional.

Como cabe esperar, la mayoría de las ejecuciones previstas se llevarán a cabo bajo la forma de avisos y recomendaciones automatizables que se ofrecerán para aquellos fenómenos en los que, para su tratamiento, se pueda prescindir de una interpretación semántica profunda. Se abordarán aquí algunos asuntos propios del repertorio de los coloquialismos y se ofrecerán varias opciones, ajustadas al criterio estilístico de la norma, para *decir lo mismo*.

15.1 Dos formas para decir lo mismo

En este epígrafe se recoge el tratamiento previsto para casos en los que, no siendo incorrecto ni agramatical el patrón identificado, cabe mejorarlo ajustando algún elemento a las recomendaciones expresas por la norma.

Se han codificado 99 patrones que abordan aquellos casos que presentan una doble opción para la expresión correcta, pero solo una es la recomendada por la norma culta. Se aborda este contenido organizándolo en dos bloques; uno dedicado a las palabras correctas de uso frecuente que tienen posibilidad de mejora, de acercarse de manera definitiva al estilo del español formal, y otro con

expresiones que del mismo modo ofrecen posibilidad de apropiarse de una variedad del español más precisa y elegante.

15.1.1 Palabras

15.1.1.1 *Superlativos*

Se han diseñado recomendaciones para algunos adjetivos que presentan dos formas de superlativo. Una está asociada a la variedad coloquial y se considera descuidada con respecto a la otra, la avalada y recomendada por la norma culta. Esta doble opción entre versiones del superlativo se encuentra en dos bloques diferenciados de adjetivos. Por un lado, la incidencia se ciñe a algunos adjetivos en los que la sílaba tónica tiene un diptongo que no presentaba en latín; *bueno, fuerte, nuevo*, etc. Se recogen los de uso más frecuente:

`@bueno@@{superl}`	`[-buenísim][+bonísim]`
`@caliente@@{superl}`	`[-calientísim][+calentísim]`
`@cierto@@{superl}`	`[-ciertísim][+certísim]`
`@diestro@@{superl}`	`[-diestrísim][+destrísim]`
`@grueso@@{superl}`	`[-gruesísim][+grosísim]`
`@fuerte@@{superl}`	`[-fuertísim][+fortísim]`
`@nuevo@@{superl}`	`[-nuevísim][+novísim]`
`@tierno@@{superl}`	`[-tiernísim][+ternísim]`
`@valiente@@{superl}`	`[-valientísim][+valentísim]`

Por otro lado, se observan pares en los que alternan la forma coloquial, que se construye sobre el aspecto que el adjetivo tiene el español actual, *amiguísimo*, con la variante más culta cuya forma deriva directamente de la forma etimológica que el adjetivo presentaba en latín; *amicísimo*.

`@amigo@@{superl}`	`[-amiguísim][+amicísim]`
`@amplio@@{superl}`	`[-amplísim][+amplicísim]`
`@áspero@@{superl}`	`[-asperísim][+aspérrim]`

@fiel@@{superl}	[-fielísim][+fidelísim]
@fríо@@{superl}	[-friísim][+frigidísim]
@negro@@{superl}	[-negrísim][+nigérrim]
@pobre@@{superl}	[-pobrísim][+paupérrim]
@pulcro@@{superl}	[-pulcrísim][+pulquérrim]
@sagrado@@{superl}	[-sagradísm][+sacratísim]

15.1.1.2 *Miscelánea*

Se codifica un pequeño repertorio de adjetivos que, pese a tener doble flexión de género, la norma recomienda emplear su versión femenina aunque el sustantivo al que complemente sea masculino; *un hombre políglota; un escudo gualda; un alumno autodidacta*.

polígloto autodidacto gualdo

@autodidacto@@{masc}	[-@@{masc}][+@@{fem}]

Otro caso codificado sirve para ofrecer la versión recomendada del adverbio *quizá:*

quizás	quizá	Aunque ambas formas son válidas, la RAE aconseja utilizar la versión escueta *quizá*.

Por otro lado, es común encontrarse, en textos propios de casi todos los registros y variedades, la expresión del concepto que se aloja en la preposición *contra,* a partir de la abreviatura *vs.* o *versus*. Aunque esta preposición propia del inglés ha sido admitida recientemente por la Academia, no se recomienda su uso en ningún caso, ni tampoco en su versión apocopada, *vs*. Se aconseja sustituirla por sus equivalentes en español *contra* o *frente a*.

Se codifica, para ejecutar esta recomendación, el siguiente bloque de patrones, que se replicarán para el tratamiento sobre la forma en cursiva:

versus	contra|frente a
vs	contra|frente a
vs.	contra|frente a

Por último, se presenta el siguiente patrón como ejemplo del funcionamiento y potencial del subprograma dedicado a desarrollar las relaciones morfoléxicas. Codificando esta opción en la base de datos se generarán de modo automático todas las posibilidades morfológicas y léxicas que este sustantivo pueda desarrollar; así se obtendrán otras formas frecuentes como *yerbabuena*.

@yerba	[-yerb][+hierb]	Relaciones morfoléxicas

15.1.2 Expresiones

15.1.2.1 *Expresiones partitivas*

Siguiendo las orientaciones normativas, las construcciones partitivas del tipo *la mayoría de, una parte de, la mitad de*, etc., deben hacerse acompañar del artículo que delimita al nombre al que acotan.Se prefieren expresiones como *La mayoría de los asistentes se fueron antes del final del concierto,* a la opción despojada del artículo ⊗*Una gran parte de gente prefiere otro sistema de gobierno,* a pesar de que la tendencia general opta por esta última opción.

Para ofrecer un tratamiento estilístico a estas expresiones se han diseñado las siguientes baterías de patrones que abordan todas las posibilidades flexivas que ofrece la casuística:

la mayor parte de @{N}@@{fem pl}	[- de][+ de las]
la mayor parte de @{N}@@{fem sg}^@@{propio}	[- de][+ de la]
la mayor parte de @{N}@@{masc pl}	[- de][+ de los]
la mayor parte de @{N}@@{masc sg}^@@{propio}	[- de][+ del]
Para construcciones partitivas del tipo la mayoría de, la mayor parte de, etc. que señalan una fracción dentro de un conjunto, la norma prefiere la expresión con el artículo o el determinante precediendo al nombre.	

Como puede advertirse, es necesario excluir del elenco de sustantivos los propios para evitar la corrección inadecuada de oraciones como *La mayor parte de Suecia es muy religiosa; Van a hacer peatonal la mitad de Madrid.*

Este conjunto de patrones se replica con las siguientes expresiones partitivas:

`el resto de` `un cuarto de` `la mitad de`
`una parte de` `la minoría de`

Para agotar todas las opciones que pueden registrarse a este propósito, se ha codificado un grupo de patrones dedicados a expresiones numerales partitivas, fraccionarias y porcentuales, que deben seguir las mismas consideraciones estilísticas que las estructuras precedentes.

`@{num}@@{card} @{num}@@{fracc} de @{N}@@{fem pl}`	`[- de ][+ de las ]`
⊗Tres cuartos de mujeres prefieren libros de aventuras.	

`@{num}@@{card} @{num}@@{fracc} de @{N}@@{fem sg}^@@{propio}`	`[- de ][+ de la ]`
⊗Un quinto de población mundial tiene un portátil.	

`@{num}@@{card} @{num}@@{fracc} de @{N}@@{masc pl}`	`[- de ][+ de los ]`
⊗Dos octavos de alumnos secundaron la huelga.	

`@{num}@@{card} @{num}@@{fracc} de @{N}@@{masc sg}^@@{propio}`	`[- de ][+ del ]`
⊗Un tercio de plástico es reciclado.	

`@{numeral}@@{card} por ciento de @{N}@@{fem pl}`	`[- de ][+ de las ]`
⊗El cuatro por ciento de páginas web son educativas.	

`@{numeral}@@{card} por ciento de @{N}@@{masc pl}`	`[- de ][+ de los ]`
⊗El quince por ciento de escritores sueñan con un premio Nobel.	

`@{numeral}@@{card} por ciento de @{N}@@{masc sg}^@@{propio}`	`[- de ][+ del ]`
⊗El cincuenta por ciento de aforo estaba vacío.	

`@{numeral}@@{card} por ciento de@{N}@@{fem sg}^@@{propio}`	`[- de ][+ de la ]`
⊗El veinte por ciento de poesía trata sobre la muerte.	

Estas últimas baterías deben codificarse junto con los patrones que recojan todas las posibilidades formales y tipográficas que puedan desarrollar estas expresiones; palabras, cifras, signos (%) y/o la combinación de ambos.

15.1.2.2 *Expresiones con* dar

Se recogen algunas expresiones desarrolladas por el verbo *dar* junto con un sustantivo al que, contraviniendo el estilo del español formal, se le hace seguir innecesariamente de un *de* cuando el siguiente elemento es un verbo; ⊗*Me da mucha alegría de verte; ⊗No le daba ningún apuro de llamar a esas horas; ⊗Nos daba mucha lástima de observar cómo la enfermedad acabó con ella*.

Se ha comprobado la inviabilidad de diseñar un patrón general que en lugar de especificar esos sustantivos se codificara con la etiqueta

`de @{N}`

debido a la existencia de contraejemplos válidos como *Me dan ganas de abrazarla, Allí dan cosas de comer; En el avión da antojo de fumar.*

`#dar @{adj}? alegría de #{vb}`	`[- de ]`

Los sustantivos que han sido objeto de codificación diferenciada son los que siguen:

`apuro`	`lástima`	`rabia`
`asco`	`pena`	`vergüenza`

15.1.2.3 *Expresiones reflexivas*

Son objeto de consideración y tratamiento algunas expresiones reflexivas del tipo ⊗*Se retrataron a ellos mismos,* que no son preferidas por la norma ante la variante con el pronombre *sí*: *Se retrataron a sí mismos.* Para subsanar esta falta de estilo se codifican los siguientes patrones:

#{vb}##{reflex} a él mismo	[-a él mismo][+a sí mismo]
⊗*Se ridiculizó al él mismo.*	

#{vb}##{reflex} a ella misma	[-a ella misma][+a sí misma]
⊗*Se lesionó a sí misma.*	

#{vb}##{reflex} a ellas mismas	[-a ellas mismas][+a sí mismas]
⊗*Se culpabilizaban a ellas mismas.*	

#{vb}##{reflex} a ellos mismos	[-a ellos mismos][+a sí mismos]
⊗*Se felicitaron a ellos mismos.*	

La categoría verbal debe ceñirse solo al paradigma de reflexivos sin enclíticos, para evitar falsos positivos, o en este caso, recomendaciones equívocas para casos legítimos como *Entrégaselo a ellos mismos.*

15.1.2.4 *Expresiones muy frecuentes*

Por último, se ofrece una miscelánea de expresiones que, con documentada frecuencia, se leen en textos de diversos registros. Todas estas expresiones, aun siendo correctas, tienen un margen de mejora o reajuste estilístico.

**Se vende pisos*	*Se venden*
se #{vb}##{3ª sg}##{pres ind} @det+? @{adj}? @{N}@@{pl}	[-##{3ª sg}] [+##{3ª pl}]
Aunque esta opción es correcta, la norma culta prefiere la variante con el verbo en plural: *Se alquilan habitaciones*; *Se arreglan todo tipo de aparatos.*	

a día de hoy	hoy día|hoy por hoy| hoy en día|en la actualidad
Aunque esta expresión está muy extendida en el habla actual, sobre todo en el registro periodístico, político y administrativo, se considera un calco innecesario del francés. En su lugar se recomiendan otras expresiones como *hoy en día; hoy por hoy; hoy día,* etc.; ⊗*A día de hoy se observan conductas ecologistas jamás pensadas.*	

a la mayor celeridad posible	con la mayor celeridad
Aunque esta expresión es correcta, la opción recomendada es con la preposición *con*: *con la mayor celeridad.* En ambos casos, no obstante, está desaconsejado el uso del adjetivo *posible* por ser redundante. ⊗*Necesito el informe a la mayor celeridad posible.*	

**Unos actores de lo más atractivos*	*Unos actores de lo más atractivo.*
de lo más @{adj}@@{pl}	[-@{adj}@@{pl}][+@{adj}@@{masc sg}]
Aunque esta expresión es correcta, la norma culta prefiere la variante en la que no hay concordancia entre el adjetivo y el nombre al que se refiere. Es preferible *Unas novias de lo más romántico* a *#Unas novias de lo más románticas*. Cuando el nombre referido es femenino y singular, se recomienda la concordancia en femenino: *Una novia de lo más romántica*, aunque la opción no concordada también es correcta, como advierte la Academia. ⊗*Disfruta de unas vacaciones de lo más románticas.*	

violencia de género	[-de género][+doméstica]|[+machista]
Esta expresión, producto de un anglicismo, da lugar a una notable imprecisión. Por este motivo, la Academia prefiere emplear expresiones como *violencia doméstica* o *machista* o *discriminación* o *violencia por razón de sexo*. ⊗*De nuevo aumenta la cuota de víctimas de violencia de género*.	

15.2 Coloquialismos

Bajo esta etiqueta adscrita al campo de REGISTRO se han codificado los casos de algunas expresiones frecuentes propias del registro coloquial que, por lo tanto, no se recomiendan en la lengua escrita en un registro neutro. Mediante recomendaciones, el programa informará sobre el contexto propio de estas expresiones y ofrecerá una medida de mejora para aplicarla sobre la incidencia detectada.

15.2.1 Inicio de frase

Son frecuentes, y desaconsejadas en la lengua escrita, expresiones que típicamente inician el discurso dejando marcas de un registro no formal:

vaya @{N} que	[- que]	⊗*Vaya hombre que está hecho.*
vaya @{adv} que	[- que]	⊗*Vaya despacio que venía.*

qué @{N} que	[- que]	⊗*Qué inútil que era.*
qué @{adv} que	[- que]	⊗*Qué rápido que llegó.*
qué @{adj} que	[- que]	⊗*Qué guapos que eran.*
@menudo @{N} que	[- que]	⊗*Menudas pintas que traía.*
@menudo @{adv} que	[- que]	⊗*Menudo lento que traía el coche.*
@menudo @{adj} que	[- que]	⊗*Menudo tonto que se ha vuelto.*

Como se observa, ha sido necesaria la codificación de más de un patrón por expresión para dar cobertura a todas las opciones cotextuales. Otro caso frecuente, sobre todo en el registro oral, es el que sigue:

primero que todo	antes que todo	Esta expresión, propia del registro coloquial, está desaconsejada en la lengua formal y escrita. Es preferible, en lugar de *primero*, el uso del adverbio *antes* con su mismo valor.
primero de todo	antes de todo	

Este patrón se codifica también con la forma *nada*; ⊗*Manuel, primero que nada disculparme por la ausencia de estos días.*

15.2.2 *de seguido*

Otras expresiones coloquiales de uso común objeto de recomendación son las siguientes:

de seguido	ininterrumpidamente| sin interrupción| a continuación	Aunque es una expresión aceptada, no es aconsejable utilizar esta locución en la lengua culta o cuidada. En su lugar se recomienda el uso de *ininterrumpidamente*, *sin interrupción*, o *a continuación.*
de seguida		
de corrido		
a seguido		

15.2.3 *Más* intensificador

Se han recogido, además, las siguientes expresiones frecuentes —sobre todo en el español de América—, junto con la opción recomendada por la norma culta:

`más nunca`	`nunca más`	*⊗No vuelvas más nunca.*
`más nadie`	`nadie más`	*⊗No invites a más nadie.*
`más nada`	`nada más`	*⊗No pidas más nada.*

15.2.4 Expresión de la duda

No se considera apropiado, en la lengua escrita o en un registro formal, el uso del término *igual* o la locución *a lo mejor* como adverbio de duda similar a *acaso*, *quizá* o *tal vez*. Para estos usos, se recomienda utilizar estos adverbios. Se presentan los siguientes patrones de aviso lingüístico para enriquecer expresiones como ⊗*Igual viene enfadado;* ⊗*A lo mejor se calma pronto*.

`igual #{vb}##{indic}`	`[-igual][+acaso]|[+ quizá]|[+tal vez]`
`a lo mejor #{vb}##{indic}`	`[-igual][+acaso]|[+ quizá]|[+tal vez]`

Capítulo 16

Tratamientos transversales

En este último capítulo se abordan fenómenos cuyos desarrollos y tratamientos totales exigen la intervención en más de un nivel de la lengua. La construcción de una expresión puede enfocarse desde diferentes perspectivas lingüísticas, el tratamiento global de los términos extranjeros, las expresiones latinas o los vulgarismos también requiere la observación de diferentes niveles. Por último, los errores derivados de un fracaso en la segmentación gráfica de términos o secuencias incide sobre la ortografía y la semántica de una oración.

Por motivos de claridad en la exposición se ha optado por agrupar estos asuntos y ofrecer aglutinadas todas las medidas que se tienen previstas para el tratamiento de ciertos fenómenos generales.

16.1Construcción de una expresión

Para subsanar ciertas incidencias que se observan en el seno de algunas expresiones frecuentes del español actual es necesaria la intervención sobre diferentes planos. Los inconvenientes que presentan estas estructuras erróneas casi lexicalizadas pueden deberse al fallo en diferentes estratos de la lengua. A partir de estos niveles, se organizará la casuística tratada en relación con la construcción de expresiones.

16.1.1 Errores en la ortografía

16.1.1.1 *Adverbios prefijados con* a-

Dentro de los casi 1 800 patrones dedicados a errores e incidencias en el plano ortográfico, cabe hacer mención a ciertos equívocos ortográficos que se registran en expresiones de uso diario en las que intervienen adverbios prefijados con *-a*; **Emprendió una huida hacia adelante; *Fue un caballero y la acompañó hasta adentro del taxi; *Solo espero encontrar nuevos aires afuera de estas paredes.* Para solventar este tipo de errores, se han estudiado y abordado dos fenómenos.

- Por un lado, se han diseñado tres bloques de patrones que evitan que los adverbios prefijados con *-a* se combinen con preposiciones que indiquen dirección como *hacia, hasta* o *para.* Este tipo de coaliciones solo son aceptables en el español de México, por lo que, si no se selecciona este perfil de usuario, la corrección sobre el texto será automática.

`hacia adentro`	`hacia dentro`
`hacia afuera`	`hacia fuera`
`hacia atrás`	`hacia tras`
`hacia adelante`	`hacia delante`

Se replican estos patrones con las preposiciones

`hasta` `para`

- Por otro lado, se han desarrollado patrones que eviten casos como **Mira adentro de mi bolso.* Siguiendo las consideraciones de la norma del español europeo, no es admisible la expresión de adverbios prefijados con *-a* seguidos de la preposición *de.* Esta restricción no se extiende al español de América, donde son construcciones legítimas y de uso extendido. Este matiz constará entre las condiciones de aplicación de este patrón.

adentro de	dentro de
afuera de	fuera de
atrás de	tras de
delante de	delante de

16.1.2 Errores gramaticales

16.1.2.1 *Rección de algunos adverbios de lugar*

A pesar de lo común de expresiones como **Lo dijo delante mío; Se posicionó cerca suya*, la construcción correcta con estos adverbios de lugar exige acompañarlos de la preposición *de* junto con el pronombre personal tónico, es decir, la serie de *mí, ti, él, ella, nosotros, vosotros, ellos*. En ningún caso se considerará aceptable la contigüidad entre estos adverbios y un pronombre posesivo. Se codifica, para solventar estos casos, la siguiente batería de patrones

delante @mío	delante de mí
delante @nuestro	delante de @nosotros
delante @tuyo	delante de ti
delante @vuestro	delante de @vosotros
delante suya	delante de ella
delante suyo	delante de él
delante @suyos	delante de @ellos

Hay que hacer notar, a propósito de estos patrones, que las variantes de tercera persona *suyo* y *suya* tienen que desdoblarse y especificarse en una entrada única para facilitar la ejecución de la corrección que presenta dos opciones según el género. No sucede con el resto de las formas cuya corrección es unívoca independientemente del género y número que presente el error.

Este batería se desarrolla homólogamente con los siguientes adverbios;

arriba	debajo	encima
abajo	detrás	lejos
cerca	enfrente	

Y con dos sintagmas preposicionales con las que suele coaparecer *en contra* y *a favor*.

16.1.2.2 De *impropio en lugar de la conjunción* que

Es frecuente, sobre todo en el registro oral, toparse con expresiones del tipo **Yo de ti, rechazaría la propuesta,* en las que hay un uso impropio de la preposición *de* sustituyendo a la conjunción requerida *que.* La sustitución de la preposición por la conjunción en este tipo de estructuras posiblemente se deba a la influencia del catalán, donde son aceptables las estructuras con la preposición. A pesar de su difusión, esta expresión de uso en el español europeo y en algunas zonas de Centroamérica está desaconsejada por las fuentes normativas, en cualquier caso, registro y variedad.

`@{pn}@@{sujeto} de @{pn},`	`[- de ][+ que @{pn}@@{pers}]`
**Yo de ti, volvería a intentarlo.*	

`@{pn}@@{sujeto} de @{N},`	`[- de ][+ que ]`
**Yo de Alba, cambiaría de asesor.*	

`@{pn}@@{sujeto} de @{det}+ @{adj}? @{N},`	`[- de ][+ que ]`
**Yo de esas tres mujeres, hubiera puesto una denuncia.*	

16.1.2.3 *Algunas expresiones recíprocas*

Para los casos de expresiones recíprocas que contienen los pronombres *uno* y *otro*, la norma recomienda que la concordancia que haya de establecerse lo haga con

género en masculino —género no marcado—, incluso en los casos mixtos en los que *unos* y *otros* se refieran a seres de diferentes sexos.

el uno {prep}^}^(de|a) la otra	[-la otra][+al otro]
⊗*Tío y sobrina se miraron el uno a la otra.*	

los unos {prep}^(de|a) las otras	[-las otras][+los otros]
⊗*Los unos querían casarse con las otras.*	

los unos {prep}^(de|a) la otra	[-la otra][+el otro]
⊗*Los unos cantaron canciones de amor a la otra.*	

Aunque se podría establecer un patrón genérico e irrestricto con la categoría {preposición}, sin hacer la exclusión de *a* y *de*, debe tenerse en cuenta que la corrección para aquellos casos en los que las preposiciones sean estas deberá estar encabezada por las contracciones *al* y *del.* Por este motivo se codifican diferenciados estos casos y se excluyen del patrón general con el fin de no interferir en la corrección general.

los unos de las otras	[-de las otras][+de los otros]
los unos de la otra	[-de la otra][+del otro]
el uno de la otra	[-de la otra][+del otro]
el uno a la otra	[-a la otra][+al otro]
los unos a la otra	[-a la otra][+al otro]
los unos a las otras	[-a las otras][+a los otros]

16.1.2.4 *Errores causados por la rección preposicional*

Se expone, a continuación, el tratamiento específico que se ha llevado a cabo para solventar algunos de los errores que ciertas formas generan por restricciones que se aplican a sus cotextos.

Abajo y *arriba*

Las formas adverbiales *abajo* y *arriba* rechazan la anteposición de la preposición *a* debido a que esta queda incluida en su forma y su significado. Si es aceptable, no obstante, la anteposición de otras preposiciones; *Iba hacia abajo cuando lo encontré; Se oían las voces desde arriba de la azotea.*

Por otro lado, estos adverbios no admiten ser seguidos por complementos preposicionales cuyo término sea *de*. En su lugar, deben utilizarse las variantes *debajo* y *encima*.

Para ofrecer un tratamiento adecuado a estas formas siguiendo las orientaciones normativas, se han diseñado los siguientes patrones.

a abajo	abajo	**Búscalo a abajo.*
a arriba	arriba	**Se dirigía a arriba de la plaza.*
abajo de	debajo	**Lo vi abajo de tu almohada.*
arriba de	encima	**Ponlo arriba de todo.*

Se codifica, además, la expresión errónea lexicalizada **de arriba a abajo,* que rechaza la preposición *a,* ya que su significado, *de principio a fin,* no necesita la inclusión de una preposición de dirección.

abajo a arriba	abajo arriba	**Me estudié todo de abajo a arriba.*
arriba a abajo	arriba abajo	**Me miró disimuladamente de arriba a abajo.*

Antes y *después*

Estos adverbios que con frecuencia se hacen seguir por la preposición *a* deben, en su lugar, acompañarse de la forma *de,* siguiendo las recomendaciones propuestas por la norma.

antes a	antes de
antes al	antes del

después a	después de
después al	después del

A propósito del régimen de estas formas, que pueden ser modificadas por otros adverbios como *mucho, bastante* y *algo*, debe hacerse un matiz; no se acepta su combinación con *más; *Ven lo más antes posible.* En su lugar se recomienda utilizar *lo antes* —por **más antes*— o *más tarde* —por **más después*—.

más antes	lo antes
más después	más tarde

Por último, con respecto al adverbio *antes,* se elabora una recomendación para evitar, siguiendo las orientaciones normativas, expresiones como **Paco llegó antes de mí; *Antes de ti, yo ya sabía cómo era.*

En lugar de este sintagma preposicional, es preferible utilizar expresiones como *antes que yo* o *delante de mí*, según sea el sentido y contexto comunicativo. Por este motivo —necesidad de información pragmática—, no se automatiza la medida correctiva y se codifica una recomendación que ayude al escritor a corregir la secuencia de modo interactivo.

antes de mí	antes que yo|delante de mí
antes de sí	antes que él|antes que ella| delante de él|delante de ella
antes de ti	antes que tú| delante de ti

Al contrario* y *al revés

Las expresiones *al* contrario y *al revés* precisan la forma *que* cuando estas van seguidas de un grupo nominal o preposicional; *Al contrario que Margarita; Al contrario que en Alemania*. Acompañadas de otros contextos, cuando le sigue toda una oración, sí es correcto utilizar la preposición; *Al contrario de lo que todos pensaban, fui capaz de hacerlo.*

al revés de {prep}	[- de][+ que]
**Al revés de en Alemania,*	

al revés de @{pn}	[- de][+ que]
**Al revés de ti,*	

al revés de @{N}	[- de][+ que]
**Al revés de Pedro,*	

al revés de @{det}+? @{adj}? @{N}	[- de][+ que]
**Al revés de aquellos silentes estudiantes,*	

Se codifica este mismo repertorio de patrones con la expresión errónea:

al contrario de

Rección de algunos sustantivos

La semántica de algunos sustantivos exige una rección preposicional, en ambos cotextos, que a menudo se ve alterada. Locuciones como *en virtud de, con relación a, poner de relieve* o *estar harto a,* se construyen sobre sustantivos como *virtud, base, relación* o *relieve,* que seleccionan nítidamente sus contornos. Se recoge, a continuación, el repertorio de algunas de estas expresiones nominales alteradas que son objeto de error. Se ha incluido, junto con estas, el adjetivo *harto* y el verbo *hartar,* que con frecuencia se presentan en estructuras erróneas.

en virtud a	en virtud de
**En virtud a aquél acuerdo se desarrolló el proyecto.*	

en relación a	con relación a|en relación con
**En relación al tema inmobiliario hay mucha letra pequeña.*	

#poner en relieve	[- en][+ de]
**No puso en relieve el tema verdaderamente importante.*	

bajo el punto de mira	en el punto de mira

**Aquel pueblo estuvo bajo el punto de mira.*

@harto a #hartar a	[- a][+ de]
**Está harto a leer libros de caballerías.*	

Se codifican, además, un par de patrones para tratar el extendido caso de incongruencia que se genera al intentar coaligar los términos *bajo* y *base*. La preposición adecuada y lógica para esta expresión es *sobre.*

bajo @{det}+ @base de	[-bajo][+sobre]
**Se aplicarán las políticas sociales bajo la base de la situación económica actual.*	

en base a	sobre la base de|con base en
**En base a tu criterio todos deberíamos callarnos.*	

16.1.2.5 *Algunos pronombres indefinidos*

La norma culta no admite expresiones como **Alguien de ellos robó las claves*; **Nadie de nosotras supo contestarle como se merecía,* en las que se combina un pronombre indefinido con un complemento preposicional cuyo término es *nosotros, vosotros* o *ustedes.* La configuración correcta de estas estructuras distributivas es la encabezada por los indefinidos *alguno* o *ninguno*.

Se presenta, a continuación, la batería que recopila la casuística que ofrece este tipo de expresiones:

alguien de nosotros	alguno de nosotros
alguien de nosotras	alguna de nosotras
alguien de vosotros	alguno de vosotros
alguien de vosotras	alguna de vosotras
alguien de ustedes	alguno de ustedes
alguien de ellos	alguno de ellos
alguien de ellas	alguna de ellas

Se codifica otro bloque homólogo para las estructuras erróneas con *nadie* que serán corregidas sustituyendo este indefinido por;

```
ninguno  ninguna
```

16.1.3 Errores de estilo

16.1.3.1 *Expresiones de tiempo: adverbios y preposiciones*

Antes* y *después

Cuando los adverbios *antes* y *después* están precedidos de sustantivos con valor temporal como *día, noche, semana*, etc., la norma prefiere prescindir de la preposición *de*. Para subsanar casos como ⊗*El secuestro se llevó a cabo la noche de antes del conflicto* se codifican patrones con la siguiente estructura:

`@semana de después`	`[-de después][+después]`
`@semana de antes`	`[-de antes][+antes]`

Se redactan pares de patrones para otros sustantivos que típicamente protagonizan estas expresiones:

```
@segundo        @año          @día
@minuto         @mes          @noche
@hora
```

A propósito de estos sustantivos, se han recogido algunas expresiones distributivas que presentan un error de rección preposicional.

A* y *por

Para la expresión se secuencias distributivas que impliquen cantidades y unidades de tiempo superiores a la hora, la norma recomienda encabezarlas con la preposición *a*; *dos veces al mes; cien euros a la semana,* etc.

Sin embargo, cuando las unidades de tiempo son menores a la hora, la preposición recomendada es *por*; *Venía a 120 km por hora; Difícilmente ganaba más de tres euros por hora.*

Para subsanar estos errores de estilo se han codificado los siguientes patrones:

por semana	a la semana
por mes	al mes
por día	al día
por año	al año

a la hora	por hora
al minuto	por minuto
al segundo	por segundo

Entre* y *contra

Por último, dentro de este epígrafe dedicado a errores de estilo en algunas expresiones temporales, se ha recogido los siguientes pares de secuencias frecuentes que presentan margen de mejora.

contra más	mientras más
contra más	cuanto más

entre más	mientras más
entre más	cuanto más

16.1.3.2 *Expresiones numerales*

Una de las muchas incidencias que se nos ofrecen a propósito de las expresiones numerales se retrata en oraciones como ⊗*Tenía treinta y tantos de años y aún no sabía hacer un huevo frito;* ⊗*Me cobraron 40 y pico de euros por la entrada,* en las

que se registra un uso afuncional de la preposición *de.* Por este motivo se desaconseja utilizar esta preposición entre el numeral compuesto y el sustantivo al que cuantifica.

Para corregir estos errores estilísticos frecuentes se han desarrollado los siguientes patrones que contemplan las dos posibilidades formales de la expresión del numeral, tanto con dígito como con término;

`\d+ y tantos de`	`[- de]`	**30 y tantos de grados*
`\d+ y pico de`	`[- de]`	**30 y pico de años*
`@{numeral}@@{card} y tantos de`	`[- de]`	**Treinta y tantos kilómetros*
`@{numeral}@@{card} y pico de`	`[- de]`	**Treinta y pico de euros*

16.1.3.3 *Expresiones de enfoque*

Es muy frecuente, sobre todo en el entorno de los medios de comunicación, encontrar expresiones del tipo ⊗*bajo la perspectiva occidental;* ⊗*bajo el planteamiento económico moderno;* o ⊗*bajo un enfoque cognitivo.* Este tipo de conectores, empleados para señalar el modo o enfoque con que se considera un asunto, deben encabezarse con las preposiciones idóneas *desde* o *según*; *desde mi punto de vista, según mi enfoque, según este planteamiento,* etc.

Para enmendar esta falta de carácter semántico que repercute en el estilo, se ha desarrollado un pequeño bloque de patrones que siguen la siguiente estructura:

`bajo @{det}+ @punto de vista`	`[-bajo][+según]|[+desde]`

A partir de este modelo se codificaron los siguientes sustantivos que pueden presentarse en estos conectores:

```
@enfoque          @planteamiento     @visión
@perspectiva      @opinión
```

16.1.3.4 *Cacofonías*

No es recomendable, según la norma de estilo, ni la escritura de dos formas contractas seguidas, ni la de galimatías formales que concentren varios homófonos en una secuencia. Estas congregaciones generan cacofonías y no se consideran estilísticamente aceptables: oraciones como ⊗*Nunca me gustó la actitud de la de la pastelería,* ⊗*Soy paisana del del tercero;* ⊗*Prefiero el sabor del vino blanco al del tinto.*

Para evitar este fenómeno fónico, se recomienda reelaborar el texto salvando estas repeticiones; *Prefiero el sabor del vino blanco antes que el del tinto*. En otras ocasiones podrá solventarse recuperando el sustantivo omitido; *Nunca me gustó la actitud de la dueña de la panadería; Soy paisana del vecino del tercero.*

Como puede intuirse, el máximo tratamiento que pueden recibir estas incidencias será a través de un aviso lingüístico que informe de esta tara estilística y ofrezca, en abstracto, una medida que derive en solución.

Los patrones que sirven de base para activar el aviso recogen las siguientes secuencias:

```
que que              las de las de        del del
los de los de        la de la de          al del
```

16.1.4 Incidencias en el plano léxico-semántico

Como es bien reconocida a lo largo de este trabajo, la limitación que supone la ausencia de interpretación léxica y semántica por parte de la máquina impide en muchos casos desarrollar y ofrecer un tratamiento correcto automatizable que ofrezca todas las garantías de corrección. Cuando los cotextos de las formas susceptibles de error no sirven de anclaje para asegurar que esa forma debe sustituirse, la tarea de ingeniería se limita al asesoramiento lingüístico. Se presentan, a continuación, algunos de los casos recogidos en el repositorio de errores. Una vez más esta exposición relacionada con asuntos léxicos no pretende ser exhaustiva en cuanto a la cantidad de patrones registrados sino ilustrativa de los procedimientos llevados a cabo para desarrollar esta línea de trabajo.

16.1.4.1 *Interferencias entre construcciones*

Se han diseñado avisos lingüísticos que ayuden al escritor a evitar las confusiones generadas como consecuencia del parecido fónico que surge entre las locuciones *cuando más* y *cuanto más.* Su uso se somete, en numerosas ocasiones, a que surjan interferencias entre estas expresiones;

@cuanto más	*Cuando más* significa *como máximo*, *como mucho*: *Su carácter era, cuando más, soportable*. Generalmente esta expresión se rodea de comas. *Cuanto más*, por su parte, indica por un lado, *con mayor motivo, más aún: Invité a todo el mundo, cuanto más a ti, y por otro, si más: Cuantas más ganas pongas, más éxito tendrás*.
cuando más	
cuando menos	*Cuando menos* significa *como mínimo: El título del libro es, cuando menos, sugerente.* Generalmente esta expresión aparece entre comas. *Cuanto menos*, por su parte, indica, *por un lado, con menos motivo, menos aún: No quería ni verla, cuanto menos ayudarla*, y por otro, *si menos: Cuanto menos dinero tenga, menos problemas.*
@cuanto menos	

16.1.4.2 *Restricciones semánticas*

Se recogen en este epígrafe, para que sirvan de referencia, algunos ejemplos de patrones desarrollados para el tratamiento semántico de algunas expresiones que se han materializado en avisos lingüísticos.

#barajar	No debe utilizarse este verbo en el sentido de *considerar, estudiar, analizar,* cuando lo que se tiene en consideración no es más de una cosa, es decir, si el complemento directo no expresa plural. No será correcta la frase **Barajo la opción de regresar.*
@inicializar	Este verbo, propio del contexto de la informática, significa *establecer los valores iniciales para la ejecución de un programa*. Fuera de este contexto, el verbo más adecuado es *iniciar*.

la mayoría

Se ha diseñado un tratamiento específico para el sustantivo *mayoría* que contempla las restricciones semánticas que impone este partitivo; el sustantivo al que ciña debe ser de naturaleza colectiva o estar en género plural.

Para subsanar estos usos no recomendados, se codifican tres patrones específicos y un aviso general que informará al escritor sobre estas limitaciones:

mayoría de @{N}@@{masc sg}^(propio)	[- la mayoría] [+ la mayor parte]
⊗*La mayoría de día he estado en casa.*	

mayoría de la? @{N}@@{fem sg}^(propio)	[- la mayoría] [+ la mayor parte]
⊗*La mayoría de la casa ha quedado limpia.*	

mayoría del @{N}@@{masc sg}^(propio)	[- la mayoría] [+ la mayor parte]
⊗*La mayoría del cuaderno estaba lleno de notas insustanciales.*	

Como se observa, el patrón se diseña contando solo con sustantivos en singular, que son los que causarían la incongruencia, y excluyendo, como es obvio, los nombres propios. Dentro de los primeros, los sustantivos en singular, deben extraerse aquellos que están presentes en la Lista 13 —sustantivos colectivos—, que por su naturaleza quedan excluidos de esta restricción.

Para copar toda la casuística posible en cuanto presencia —recomendada— o ausencia —no recomendada— del artículo que sigue a la preposición, ha sido necesario, como se observa, triplicar el patrón abarcando las tres configuraciones previsibles.

Los casos que queden sin tratar por esta tripla de patrones podrán ser identificados por el programa y activarán el siguiente aviso lingüístico:

> *Mayoría* suele llevar un complemento encabezado por *de* al que le sigue un sustantivo que, por las restricciones que *mayoría* impone, debe ser colectivo; población, fauna, cubertería, etc., o un sustantivo en plural. No debe emplearse esta forma seguida de sustantivos no numerables como en ⊗*la mayoría del tiempo*, o en singular ⊗*la mayoría del queso*. En estos casos, se recomienda utilizar la fórmula *la mayor parte de*.

Junto a estos patrones pueden incluirse además de ciertos casos residuales, otros de su misma naturaleza que advierten sobre asuntos semánticos de expresiones que suelen suscitar dudas y errores. Por motivos de organización se

ha optado por incluirlos en el repertorio de los neologismos, las expresiones latinas, los términos casi homófonos o los mecanismos de concordancia.

16.2 Extranjerismos

Se consideran extranjerismos, en el entorno de este trabajo, a los términos o expresiones que una lengua toma de otra, ya sea para llenar un vacío semántico —tuit— o como alternativa a otras expresiones ya existentes en la lengua de destino —*parking*—.

En otros casos, las palabras adoptadas no suplen ninguna carencia léxica, pero se han ido infiltrando, a lo largo de los años, en el repertorio de la lengua hasta que finalmente son admitidos por su uso habitual —*ticket*—.

Un precepto que sostiene la norma de nuestra lengua para el uso y escritura de los extranjerismos, al margen de la motivación de su uso, es que la incorporación responda, en lo posible, a nuevas necesidades expresivas. En estos casos la Academia admite su inclusión siempre que se haga *de forma ordenada y unitaria*, acomodando estos términos a los *rasgos gráficos y morfológicos propios del español.* (NGLE § 3.4c y ss.; DPD, *extranjerismo*).

Como en el caso de otros bloques temáticos que se tratan en este trabajo latinismos, siglas, abreviaturas, redundancias, etc., el asunto de los extranjerismos puede afectar, y de hecho afecta, a varios niveles de la lengua. Mientras que la inclusión de nuevos términos atañe, en primera instancia, al nivel tipográfico y ortográfico y complementariamente a los niveles léxico-semántico y estilístico, los procesos de adaptación a nuestra lengua tendrán que ver con operaciones y reglas de los que se encarga la morfología.

El programa ofrecerá tratamiento para aquellas palabras o expresiones que, aun siendo ajenas a nuestra lengua, participan activamente en sus manifestaciones, con independencia de los registros, variedades o contextos de uso. El diseño de patrones y soluciones se ha inspirado en las siguientes directrices normativas;

I. En términos generales la norma prefiere, siempre que sea posible, evitar el extranjerismo innecesario y optar por la versión española del concepto.
II. En el caso de optar por la versión extranjera, la norma propone un ajuste del término, esto es, una adaptación a los rasgos gráficos y morfológicos de nuestra lengua.
III. Cuando el objeto de tratamiento sea una expresión derivada estructural o léxicamente de una expresión de otra lengua se ofrecerá, siguiendo la norma y estilística de nuestra lengua, la secuencia en su versión plenamente española: ⊗*asuntos a tratar* → *asuntos para/por tratar.*

Para el abordaje de estos fenómenos relacionados con los términos foráneos, el programa se apoya en 161 patrones recogidos bajo tres etiquetas de error y una de recomendación que, en virtud del sistema de listas que hemos previsto, dan cobertura y tratamiento a cientos de errores relacionados con los extranjerismos.

Para su exposición, se ha optado por presentar la masa de patrones organizada según el nivel lingüístico sobre el que se ejecuta el tratamiento.

16.2.1 Léxico: extranjerismos con versión española

Se ha recogido un repertorio creciente de extranjerismos crudos que se registran con frecuencia en los textos escritos en español. Su inclusión, en estos casos, no se justifica por colmar carencias léxicas ni semánticas que limiten la expresión. Su incorporación en nuestro repertorio es, al menos en términos léxicos, absolutamente innecesaria.

Como ya se enunció, la norma de nuestra lengua recomienda el uso de términos propios, en lugar de ajenos, para designar conceptos para los que tiene provistos términos. Así se prefieren las formas *patrocinio* o *patrocinador* a la adaptación *espónsor* o al anglicismo *sponsor.* Respetando esta premisa se han codificado 95 patrones que contienen voces extranjeras consideradas erróneas —por innecesarias— y ofrecen como alternativa su réplica en español.

En algunos casos nuestra lengua dispone de diversas variantes para evitar un mismo término extranjero. Es el caso de *niñera* o *canguro* en lugar de *baby sitter* o *anuncio* o *cuña publicitaria* para *spot*. Con respecto a estas incidencias, que pueden resolverse ejecutando diferentes opciones, el programa ofrecerá al usuario todas las variantes disponibles listadas según su peso de frecuencia que quedará establecido a partir de los datos que ofrece el Corpus TIP.

Cabe señalar, por último, que para dar cobertura a todas las posibilidades orto/gráficas que el término extranjero pueda ofrecer, en algunos casos ha sido necesaria la codificación de varios patrones para el tratamiento de un mismo concepto. Es el caso de *feedback*, *feed back* y *feed-back*, que requiere tres entradas diferentes para adoptar un mismo tratamiento.

Se ofrecen, a continuación, los patrones codificados con el extranjerismo crudo junto con la versión española que ofrece el programa;

link	vínculo|+enlace
links	vínculos|enlaces
share	cuota de audiencia
copyright	derechos de autor
copyright	derechos de autor
baby-sitter	niñera|canguro
best seller	superventas
best-seller	superventas
blue jeans	vaqueros
coach	entrenador|preparador
coaching	entrenamiento|preparación
coachs	entrenadores|preparadores
display	demostración| pantalla de visualización
displays	demostraciones| pantallas de visualización
espónsor	patrocinio
espónsors	patrocinios
esponsors	patrocinios
fair play	juego limpio
fast food	comida rápida
feed back	retroalimentación|retroacción

feedback	retroalimentación|retroacción
feed-back	retroalimentación|retroacción
finger	pasarela
grill	parrilla
hacker	pirata informático
hackers	piratas informáticos
hall	vestíbulo|recibidor|entrada
halls	vestíbulos|recibidores|entradas
handicap	desventaja|obstáculo| impedimento|discapacidad
hándicap	desventaja|obstáculo| impedimento|discapacidad
handling	servicios de tierra
hardware	equipo informático|componentes
hit	éxito
hits	éxitos
hobbies	aficiones|pasatiempos
hobby	afición|pasatiempo
holding	grupo empresarial
holligan	hincha violento
holligans	hinchas violentos
jet lag	desfase horario
lifting	estiramiento
light	bajo en calorías|ligero
lights	bajos en calorías|ligeros
lobbies	grupos de presión
lobby	grupo de presión
mailing	buzoneo
match	partido
mobbing	acoso
nursery	sala de cunas
off the record	confidencialemente| extraoficialmente|privadamente
overbooking	sobreventa|sobrecontratación
parking	aparcamiento
parkings	aparcamientos
password	contraseña
photo finish	foto de llegada
play back	pregrabado|sonido pregrabado
play off	eliminatoria

playback	pregrabado|sonido pregrabado
play-back	pregrabado|sonido pregrabado
playoff	eliminatoria
play-off	eliminatoria
prime time	horario estelar
realities	programas de telerrealidad
reality	programa de telerrealidad
reality show	programa de telerrealidad
reality shows	programas de telerrealidad
revival	resurgimiento|regreso
roulotte	caravana
royalty	regalía|canon|derechos
sex symbol	símbolo sexual
shopping	compras
short	pantalón corto
shorts	pantalones cortos
show	espectáculo
showman	animador|presentador
single	disco sencillo|soltero
skin	cabeza rapada
skinhead	cabeza rapada
skinheads	cabezas rapadas
skins	cabezas rapadas
snack bar	cafetería
software	programas|aplicaciones
speaker	altavoz|animador|locutor
speech	discurso
sport	deporte|informal
spot	anuncio|cuña publicitaria
spots	anuncios|cuñas publicitarias
stock	existencias
tour	viaje|gira|ruta turística
tours	viajes|giras|rutas turísticas

#esponsorizar	[-#esponsorizar][+#patrocinar]
#linkar	[-#linkar][+#enlazar]|[+#vincular]
#linkear	[-#linkear][+#enlazar]|[+#vincular]

La última batería expuesta dedicada a formas verbales necesita, para su aplicación, la creación previa de estos verbos erróneos y sus conjugaciones con el fin de garantizar la cobertura del patrón sobre toda la casuística.

16.2.2 Ortografía: extranjerismos adaptables al español

Este bloque, que se basa mayoritariamente en la intervención sobre el plano ortográfico, recoge términos extranjeros —sustantivos y adjetivos— que se han infiltrado en nuestro repertorio habitual por su amplia aceptación y tradición. La norma, no obstante, recomienda darles un aspecto de léxico español adaptando sus grafías, reglas morfológicas y normas de acentuación a las propias de nuestra lengua. Así para términos como *cocktail* o *express* el programa ofrecerá las versiones normativas *exprés* y *cóctel*. Para esta última, además, se ofrecerá la flexión del plural *cócteles,* siguiendo las reglas morfológicas de construcción de plurales en español.

A continuación, se ofrece una tabla con los 23 patrones que abordan estas adaptaciones partiendo de la forma foránea:

cassette	casete
cassettes	casetes
cocktail	cóctel
cocktails	cócteles
express	exprés
parking	parquin
parkings	párquines
spray	esprái
sprays	espráis
stand	estand
stands	estands
ticket	tique
tickets	tiques
troll	trol
trolls	troles
tweet	tuit
tweets	tuits
yankee	yanqui

yankees	yanquis
yanki	yanqui
yankis	yanquis
yogurt	yogur
yogurts	yogures

16.2.3 Tipografía: extranjerismos y locuciones latinas

Siguiendo las recomendaciones normativas, los extranjerismos crudos que no se han adaptado a nuestra lengua deben escribirse con letra cursiva para marcar gráficamente que el término no forma parte del repertorio del español. Dentro de este elenco deben incluirse, como expresiones foráneas que son, las locuciones y términos de origen latino, que habrán de someterse al mismo estilo tipográfico.

La incidencia que desarrolla esta instrucción recoge el siguiente par de patrones de error:

{Lista 5} *ranking*	*ranking*	Los extranjerismos crudos que no se han adaptado a nuestra lengua se escriben en cursiva para marcar gráficamente la diferencia entre esta forma y el resto de palabras en español: *ranking*, *lifting*, *coach*, etc.
{Lista 12} *ad líbitum*	*ad líbitum*	Las formas y locuciones latinas o griegas que no están plenamente integradas en nuestra lengua deben escribirse en cursiva siguiendo la misma norma que los extranjerismos no adaptados: *ad líbitum*, *a posteriori*, *ad calendas graecas*, etc.

Como puede advertirse, para el desarrollo de estos patrones ha sido necesario elaborar dos listas. Una, la número 5, recoge los extranjerismos crudos ausentes en el Lexicón TIP que no han recibido tratamiento previo bien mediante sinónimos, bien mediante adaptación ortográfica o morfológica. La otra, número 12, recoge términos y locuciones latinas. Ambas se someterán al tratamiento tipográfico previsto para esta incidencia, que cambiará el estilo de letra, de redonda a cursiva; *online, hashtag, outfit, sushi, ab initio,* etc.

16.2.4 Estilo: expresiones tomadas de estructuras foráneas

Es cada vez más frecuente el uso y escritura de expresiones que siguen patrones estructurales de otras lenguas, especialmente de la inglesa o la francesa. Aunque secuencias como ⊗*la no asistencia* o ⊗*el no cumplimiento* son gramaticalmente correctas en nuestra lengua, la norma prefiere el uso de expresiones y estructuras más naturales para el español. En estos casos concretos, la anteposición de la partícula de polaridad negativa *no* a un nombre abstracto no es recomendable, y delata un uso anglicado innecesario en nuestra lengua. Es preferible, en su lugar, la incorporación del antónimo del sustantivo abstracto en cuestión; *la ausencia* y *el incumplimiento.*

Se registra, a propósito de estas estructuras, la siguiente batería de pares de patrones —singular y plural—[100] que contiene expresiones gramaticales aunque desaconsejadas —por estar tomadas directamente del inglés—, junto con la opción recomendada por la estilística de nuestra lengua.

el no bebedor	el abstemio
los no bebedores	los abstemios
el no cumplimiento	el incumplimiento| la inobservancia
los no cumplimientos	los incumplimientos| las inobservancias
la no aprobación	el rechazo| la desestimación
las no aprobaciones	los rechazos| las desestimaciones
la no asistencia	la inasistencia| la falta de colaboración
las no asistencias	las inasistencias| las faltas de colaboración
la no comparecencia	la incomparecencia| la ausencia
las no comparecencias	las incomparecencias| las ausencias
la no conformidad	la disconformidad| la discrepancia| el desacuerdo
las no conformidades	las disconformidades| las discrepancias|

[100] Para este caso se han hecho expresas en patrones diferenciados las dos posibilidades flexivas para facilitarle al sistema la tarea de corrección.

	los desacuerdos
la no existencia	la carencia| la inexistencia| la omisión
las no existencias	las carencias| las inexistencias| las omisiones
la no intervención	la pasividad| la abstención
las no intervenciones	las pasividades| las abstenciones
la no proliferación	la escasez| la reducción
las no proliferaciones	las escaseces| las reducciones
la no protección	la desprotección
las no protecciones	las desprotecciones
la no renovación	la cancelación| la clausura
las no renovaciones	las cancelaciones| las clausuras
no cumplimiento	incumplimiento
no cumplimientos	incumplimientos
obra de no ficción	obra documental| obra realista
obras de no ficción	obras documentales| obras realistas

En este mismo bloque se inscriben otros casos de estructuras desaconsejadas por la norma por estar tomadas de configuraciones del francés que resultan ajenas o poco naturales en nuestra lengua, aunque es cada vez más común, en el registro escrito, el hallazgo de expresiones como *camisa ⊗a rayas, ⊗cocina a gas* o *⊗barco a vapor.* Estas configuraciones compuestas por «N *a* N» están desaconsejadas por la norma, que prefiere la sustitución de la preposición *a* por *de,* empleada en español para introducir el complemento que expresa el modo o medio por el que funciona un determinado objeto[101].

Para atajar estos casos, se codifican los siguientes patrones que recogen las expresiones afrancesadas más frecuentes:

[101] No obstante, está aceptado por la norma el uso de la preposición *a* para introducir los complementos de sustantivos deverbales derivados de verbos de acción: *pintura al óleo, grabado al fuego* o *encuadernado a canutillo.*

@{N} a cuadros	[-a cuadros][+de cuadros]
@{N} a gas	[-a gas][+de gas]
@{N} a presión	[-a presión][+de presión]
@{N} a rayas	[-a rayas][+de rayas]
@{N} a tiras	[-a tiras][+de tiras]
@{N} a vapor	[-a vapor][+de vapor]
@{N} a pedales	[-a pedales][+de pedales]

Por otro lado, y a pesar de que la norma ha censurado en sus últimas publicaciones la expresión *un cierto* frente a simplemente *cierto* (NGLE § 15.3ñ), es muy frecuente encontrarse estructuras en la lengua escrita con este uso galicado. Para evitarlos se codifican las siguientes recomendaciones:

un cierto	cierto
una cierta	cierta
unos ciertos	ciertos
unas ciertas	ciertas

Este bloque incluye, además, un patrón que ofrece tratamiento a secuencias muy frecuentes del tipo ⊗*temas a debatir,* ⊗*proyectos a desarrollar,* ⊗*ideas a tratar,* etc. A pesar de estar muy extendidas en el lenguaje periodístico y administrativo, estas construcciones derivadas del francés cuya estructura es «N *a* inf» carecen de prestigio en la norma culta del español actual. Es preferible, según sus disposiciones, sustituir la preposición *a* por las preposiciones *por* o *para* o incluir en lugar del sintagma preposicional, una oración de relativo; *temas por tratar, temas para tratar* o *temas que hay que tratar.*

@{N} a {inf}	[- a][+ para]|[por]

Por último, se diseña un aviso orientado a evitar expresiones como ⊗*El equipo viene de ganar la medalla la semana pasada*.

#venir de {inf}	[-#venir de][+#acabar de]

La expresión «*venir de* + inf» en oraciones como ⊗*Barack Obama, que viene de ser Presidente de Estado Unidos, aterrizó en Moscú con ánimo conciliador,* se considera un galicismo incorrecto, muy habitual en los medios de comunicación. Cuando la expresión no es sinónima de *acabar de* y el verbo *venir* ha perdido su significado de movimiento prefiere evitarse esta expresión.

16.3 Expresiones y términos latinos

El repertorio de latinismos[102], entendidos estos como términos o expresiones prestados por esta lengua, recibe por parte del programa diversos tratamientos por lo que los patrones que se le han dedicado intervienen sobre diversos errores e incidencias que surgen a partir de su uso en la lengua escrita. Por este motivo se incluye este epígrafe dentro de este capítulo de TRATAMIENTOS TRANSVERSALES, por afectar a varios niveles de la lengua y tener diversas líneas de actuación.

16.3.1 Ortografía y expresión correcta

Con una etiqueta de error propia se aborda un pequeño grupo de expresiones que, por su frecuencia avasalladora, obliga a desarrollar un tratamiento específico de carácter ortográfico con el fin de asegurar la escritura correcta de estas formas.

Además de los equívocos en la ortografía, es común que el error en algunos de estos casos surja a partir de la inclusión indebida de una preposición. Tal es el caso de expresiones como **de corpore in sepulto*, **a grosso modo*, **de motu propio*. A propósito de esta última locución *motu proprio* cabe señalar que ha urgido la

[102] Este repertorio incluye un conjunto reducido de helenismos.

codificación de 3 patrones para abarcar toda la casuística errónea que surge de la combinatoria de los errores de diferentes niveles.

de motu propio	motu proprio
de motu proprio	motu proprio
motu propio	motu proprio

Otros de los casos tratados,

a groso modo	grosso modo
a grosso modo	grosso modo
groso modo	grosso modo
de corpore insepulto	corpore insepulto
de ipso facto	ipso facto
modus operandis	modus operandi
modus vivendis	modus vivendi
mutatis mutandi	mutatis mutandis
pecata minuta	peccata minuta
status quo	statu quo
strictu sensu	Stricto
subjudice	sub iudice
voz populi	vox populi

16.3.2 Morfología: el plural

Se han compilado, además, aquellos términos latinos de uso frecuente que suelen causar dudas y errores en la escritura de sus formas en plural. Se han seguido, de manera meticulosa, todas las orientaciones, excepciones y adaptaciones a nuestra lengua que propone la norma académica para la formación de plurales de estas formas frecuentes. En la glosa lingüística que se le ofrece al usuario se aportará información relativa a las reglas de formación de estos plurales.

Como en los casos tratados en el epígrafe precedente, para el tratamiento de algunas formas ha sido necesario codificar varios patrones para dar cobertura a todas las opciones que puedan conllevar error —con y sin tilde, con marcas de plural en uno u otro término en formas compuestas, etc.—.

Como se observa, en algunos casos ha sido necesario codificar parte del cotexto izquierdo para cercar la búsqueda y asegurar que el término que debe tratarse efectivamente pretendía estar en plural.

@{det}@@{pl} accesit	[-accesit][+accésits]
@{det}@@{pl} accésit	[-accésit][+accésits]
@{det}@@{pl} álter egos	[-álter egos][+álter ego]
@{det}@@{pl} álters egos	[-álters egos][+álter ego]
@{det}@@{pl} creterium	[-creterium][+cretériums]
@{det}@@{pl} cretérium	[-cretérium][+cretériums]
@{det}@@{pl} deliriums trémens	[-deliriums trémens] [+delirium trémens]
@{det}@@{pl} desideratum	[-desideratum][+desiderátums]
@{det}@@{pl} desiderátum	[-desiderátum][+desiderátums]
@{det}@@{pl} habitat	[-habitat][+hábitats]
@{det}@@{pl} hábitat	[-hábitat][+hábitats]
@{det}@@{pl} magnificat	[-magnificat][+magníficats]
@{det}@@{pl} magníficat	[-magníficat][+magníficats]
@{det}@@{pl} mea culpas	[-mea culpas][+mea culpa]
@{det}@@{pl} meas culpas	[-meas culpas][+mea culpa]
@{det}@@{pl} medium	[-medium][+médiums]
@{det}@@{pl} médium	[-médium][+médiums]
@{det}@@{pl} modus operandis	[-modus operandis][+modus operandi]
@{det}@@{pl} modus vivendis	[-modus vivendis][+modus vivendi]
@{det}@@{pl} placet	[-placet][+plácets]
@{det}@@{pl} plácet	[-plácet][+plácets]
@{det}@@{pl} quorum	[-quorum][+quórums]
@{det}@@{pl} quórum	[-quórum][+quórums]
@{det}@@{pl} requiem	[-requiem][+réquiems]
@{det}@@{pl} réquiem	[-réquiem][+réquiems]
@{det}@@{pl} superavit	[-superavit][+superávits]
@{det}@@{pl} superávit	[-superávit][+superávits]
@{det}@@{pl} tacet	[-tacet][+tácets]
@{det}@@{pl} tácet	[-tácet][+tácets]
@{det}@@{pl} tandem	[-tandem][+tándems]
@{det}@@{pl} tándem	[-tándem][+tándems]
@{det}@@{pl} ultimatum	[-ultimatum][+ultimátums]
@{det}@@{pl} ultimátum	[-ultimátum][+ultimátums]

@{det}@@{pl} vademecum	[-vademecum][+vademécums]
@{det}@@{pl} vademécum	[-vademécum][+vademécums]
magisters	magísteres
magísters	magísteres
memoranda	memorandos

corpora	corpus
córpora	corpus
@{det}@@{pl} currículum	[-currículum][+currículos]
@{det}@@{pl} data	[- data][+ datos]
currícula	currículos
curricula	currículos
currículums	currículos
forums	foros
fórums	foros
juniors	júniores
júniors	júniores
{det}@@{pl} media	[-media][+medios]
podiums	podios
pódiums	podios
referenda	referendos
seniors	séniores
séniors	seniores

En esta última batería se recogen aquellos casos de plurales vacilantes que suelen adoptar erróneamente la vocal *a* para su formación; *referenda, data, currícula,* etc. A pesar de que este rasgo morfológico era el adecuado para la formación de plurales de formas neutras en latín, la norma desaconseja esta forma de desarrollar el plural, tan habitual en el mundo anglófono, por ser ajena a las reglas de nuestro sistema. En el caso en el que, aún desaconsejada, el escritor persista en el uso de esta forma habrá que marcarla tipográficamente con cursiva como el resto de los extranjerismos.

16.3.3 La tipografía

Las formas latinas que no están plenamente integradas en nuestra lengua reciben el mismo tratamiento tipográfico que el resto de los extranjerismos. Su escritura debe ser, por lo tanto, en letra cursiva con el fin de marcar gráficamente la diferencia con respecto al léxico propio del español.

Este patrón se aplicará en un segundo o tercer rastreo del texto y tras haber subsanado los errores previos que afectan a la ortografía, a la morfología o a ambas. Solo una vez obtenida la forma escrita correctamente, se procederá a su tratamiento tipográfico. Para llevar a cabo esta instrucción se ha codificado el siguiente patrón que hace uso del sistema de listas.

{Lista 12}	Las formas latinas, así como las locuciones, que no están plenamente integradas en nuestra lengua se escriben en cursiva como sucede con el resto de los extranjerismos: *ad líbitum*, *ipso facto*, *apud*.

Cabe hacer una excepción a esta regla propuesta por las normas de estilo:

sic	[-sic][+'[sic']][103]	El adverbio latino *sic* significa *así*, y se utiliza para aclarar que la palabra o enunciado que le precede se recoge tal y como se escribió o pronunció, y no se trata, por lo tanto, de un error de transcripción. Es preferible utilizar esta forma en redonda y entre corchetes: [sic].

En el Anexo 15 se ofrece la Lista 12 que recoge el listado de los términos y expresiones latinas que serán identificadas por PatErr.

16.4Segmentación y unificación

La segmentación gráfica de un término en varias unidades o su unificación en una sola forma tiene implicaciones en el significado, la ortografía, el estilo o la

[103] Los caracteres precedidos por una comilla simple deben interpretarse como un carácter del texto y no como una marca de codificación.

gramaticalidad. Se abordarán en este epígrafe aquellos casos en los que el uso de una u otra grafía —compacta o segmentada— sea relevante en términos léxicos, ortográficos, estilísticos o gramaticales.

Para ofrecer el contenido que se ha desarrollado y tratado a propósito de este fenómeno de segmentación gráfica, se expone, de entre todas las posibilidades taxonómicas que surgen, apoyado en dos bloques; uno asume los casos en los que la operación de segmentación o unificación no implica un cambio de significado en la expresión, el otro se hará cargo de los casos que resten en los que el significado —aunque no necesariamente la gramaticalidad— permanecen inalterados.

16.4.1 Sin cambio de significado

16.4.1.1 *Términos para los que se recomienda la grafía univerbal*

Se exponen, a continuación, los patrones para los que se ha desarrollado un tratamiento de unificación de formas siguiendo la recomendación normativa que manifiesta su preferencia por la grafía univerbal. Estos patrones serán recomendaciones que ofrecen opciones que pueden ser automatizadas.

en hora buena	enhorabuena	La norma recomienda la escritura de esta expresión en una sola forma.

Adverbios

a posta	boca abajo	de prisa	entre tanto
a prisa	alrededor	en frente	sobre manera
así/asi mismo	boca arriba	en seguida	

Sustantivos

alta mar	medio ambiente	puerco espín	sobre mesa
mal humor	pavo real	quinta esencia	

En todos los casos la medida correctiva es la misma; la unificación del contenido del patrón en un solo término.

Por otro lado, se desarrollan avisos lingüísticos para ciertos grupos nominales en los que, cuando sus miembros actúan como una unidad conceptual, su grafía se prefiere compacta; ⊗*La Noche buena pasada fue penosa; Las ostras son privativas en Nochebuena.* Cuando la expresión se refiere a dos unidades conceptuales, la grafía segmentada será la correcta; *La noche buena fue aquella que pasamos a las afueras de Londres.* Ante la limitación de no poder interpretar el sentido del grupo nominal, en el aviso que se activa ante el reconocimiento de estos patrones se ofrecerá información sobre las dos posibilidades.

`mal @educado`	Cuando se hace referencia a una cualidad, la de rudo o descortés, y no a la forma en que se ha educado alguien, *mal educado*, se prefiere la forma compacta *maleducado*. Generalmente para el primer caso se utilizará el verbo *ser*, mientras que el segundo suele aparecer con *estar*.
`media noche` `medio día`	Cuando se refiere a la hora opuesta del mediodía se prefiere la grafía compacta: *medianoche*, mientras que cuando hablamos de una porción de la noche debe ir segmentado: *media noche*.

`@padre`
`@nuestro@@{masc}`
`@ave @maría@@{fem}`
	Cuando esta palabra hace referencia a la plegaria, es preferible utilizar la forma compacta: *padrenuestro*. El plural de estas palabras no se hace sobre ambos constituyentes; se debe tratar como una forma univerbal: *padrenuestros*.

`@noche @buena@@{fem}` `@noche @vieja@@{fem}`	Cuando se refiere a la noche víspera de Navidad, se prefiere la grafía compacta y con la inicial mayúscula: *Nochebuena*. El plural de estas grafías univerbales, se hará de forma regular: *Nochebuenas.*

16.4.1.2 *Términos para los que se recomienda la grafía segmentada*

Dentro de este grupo de términos cuya grafía segmentada se prefiere en la norma culta a la versión compacta se ha codificado un pequeño repertorio que subsana un par expresiones numerales ⊗*treintaipico*; ⊗*cuarentaitantos*, etc. En estos casos debe segmentarse la expresión de modo que *y pico*; *y tantos*, sean dos formas independientes.

\w+itantos de	[-itantos de][+y tantos]
\w+itantos	[-itantos][+y tantos]
\w+ipico de	[-ipico de][+y pico]
\w+ipico	[-ipico][+y pico]
\w+ipocos de	[-ipocos de][+y pocos]
\w+ipocos	[-ipocos][+y pocos]

Se excluyen de estos patrones las formas compactas encabezadas por *veinti-*, por constituir el único caso univerbal aceptado en el DRAE.

Como se observa, se codifican, además, patrones de error con la preposición *de*, que con frecuencia acompaña indebidamente a estas expresiones.

16.4.2 Con cambio de significado

Complementariamente, la lengua ofrece otros casos en los que cada forma de grafía —compacta o segmentada— es relevante y excluyente en términos léxico-semánticos; la escritura de una u otra implica un concepto diferente: *aparte/a parte; a cerca/acerca*, etc.

Como cabe imaginar, el tratamiento que se ofrecerá para estos casos se materializa en recomendaciones y avisos lingüísticos que ofrecen alternativas e información sobre el significado de las formas según estén o no adheridas. En algunos casos será el escritor quien de forma interactiva haga la intervención sobre el texto. Se exponen algunas formas y expresiones de uso frecuente que, aunque suelen escribirse segmentadamente, provocan vacilaciones por la existencia en el diccionario de la forma compacta.

a parte	Cuando esta expresión es sinónimo de *además de*, se escribe unida: *Compró vino aparte del champán.* Para referirnos a una parte de, la grafía se segmenta en tres palabras: *El virus se extendió a parte de la sociedad civil.*
aparte de	

a cerca de	Cuando esta expresión es sinónimo de *sobre*, *en relación* o *referencia a*,

acerca de	se escribe compactada: *Hablamos acerca del accidente que sufrió*. Para referirnos a *un número aproximado de*, se escribe segmentada: *El asunto afecta a cerca de mil personas.*

En los siguientes casos ha sido posible acotar las expresiones con cotexto para facilitar la identificación del patrón en el contexto preciso en el que en realidad se busca, en el que el riesgo de error es más amplio y la activación del aviso o recomendación es pertinente. De este modo, ciñendo los elementos portadores del error/incidencia, se evita que el programa responda innecesariamente ofreciendo información en contextos en los que claramente no son necesarios.

{conj} de más	[-de más][+demás]
la de más las de más lo de más los de más	[-de más][+demás]
Para hacer referencia a *lo otro*, *lo restante*, la grafía se compacta: *La historia es así, lo demás son cuentos*. Escrito separado, el significado es en exceso: *Había gente de más*.	

@{det} @{adj}? en torno	[-en torno][+entorno]
{prep} en torno	[-en torno][+entorno]
@{vb} entorno	[-entorno][+en torno]
Escrito de forma segmentada, en torno, es sinónimo de la locución alrededor de, acerca de, aproximadamente: Estábamos en torno a 200 personas. Escrito de forma compacta, *entorno*, se hace referencia al ambiente: *Vivíamos en un entorno inigualable.*	

Esta última tripla de patrones —*entorno*— se replica con los siguientes pares de términos y grupos preposicionales cuyo enlace es *sin:*

sin sentido	sinsentido
sin hueso	sinhueso
sin tierra	sintierra
sin fin	sinfín

sin sabor	sinsabor
sin vivir	sinvivir
sin razón	sinrazón
sin número	sinnúmero

sin techo	sintecho
sin sustancia	sinsustancia

sin vergüenza	sinvergüenza
sin papeles	sinpapeles

Se suman, con las mismas estructuras, las siguientes formas:

contra reloj	contrarreloj	ex abrupto	exabrupto
sobre todo	sobretodo	mal entendido	malentendido

Se ofrece, a continuación, el trabajo llevado a cabo para dar cobertura a dos de los casos de segmentación que con mayor frecuencia causan errores ortográficos; la (in)diferencia entre pares como *sino; si no, y porque; por qué.*

16.4.3 *Por qué*; *porque*; *por que*

Uno de los repertorios de homófonos que suscitan más dudas y son causa recurrente de errores ortográficos es el que recoge las configuraciones posibles que pueden adoptar las formas *por* y *que*.

Conviene, como en otros epígrafes consagrados a la homofonía, hacer un estudio de los entornos típicos que acompañan a estas formas. Para ello será necesario hacer una radiografía gramatical previa que aporte la información categorial necesaria para afrontar el escrutinio de los cotextos.

A partir de este estudio, y para afrontar la corrección de estas partículas, se han habilitado tres incidencias; dos correspondientes a errores y un aviso lingüístico que se activará toda vez que el programa identifique alguna de estas formas según las configuraciones que han sido codificadas.

16.4.3.1 *Por qué*

Por qué es una forma compuesta por la preposición *por* y el pronombre interrogativo o exclamativo —tónico— *qué*. Su uso se concentra en los inicios de las oraciones interrogativas y exclamativas, tanto directas como indirectas, en las que

generalmente asume la posición introductoria; *¿Por qué no baja el precio de la vivienda?*; *No reveló por qué se había ido de aquella manera tan airada*.

Como en otras ocasiones, la ausencia de programas que acometan análisis de alto nivel tanto en el plano sintáctico, como en el de la interpretación semántica imposibilitan un tratamiento global de toda la casuística que ofrece esta forma. Estas restricciones son aún más limitativas en aquellos casos en los que la oración interrogativa o exclamativa se presente de forma indirecta.

En cualquier caso, es posible el desarrollo de algunos patrones para las estructuras directas que ofrezcan garantías de que su aplicación automática dará lugar a secuencias correctas:

¿Porqué	¿Por qué	**¿Porqué me evitaste en el cine?*
¿Porque	¿Por qué	**¿Por que me evitaste en el cine?*
¿Por que	¿Por qué	**¿Porque me evitaste en el cine?*
¡Por que	¡Por qué	**¡Por que motivos más absurdos!*
¡Porque	¡Por qué	**¡Por que motivos más absurdos!*
¡Porqué	¡Por qué	**¡Porqué motivos más absurdos!*

Cada patrón recogido en esta batería ofrece una posible configuración errónea —grafía compacta o falta de tonicidad— que da cobertura a los equívocos más comunes cuando estas formas se sitúan en su posición típica —introductoria— y bajo su versión de interrogativa o exclamativa directa.

Con independencia del tratamiento que el programa lleve a cabo con estas secuencias, el usuario recibirá en forma de glosa la siguiente información de uso:

> Para introducir preguntas y exclamaciones, tanto directas: *¿Por qué te vas?*, como indirectas: *Me preguntó por qué te ibas*, la forma correcta es la grafía segmentada compuesta por la preposición *por* y el interrogativo tónico *qué*.
>
> En el caso de las interrogativas, la estructura *por qué* admite que añadamos la palabra *razón* o *causa*, que es inadmisible siguiendo a la grafía compacta: *¿Por qué causa te fuiste?*

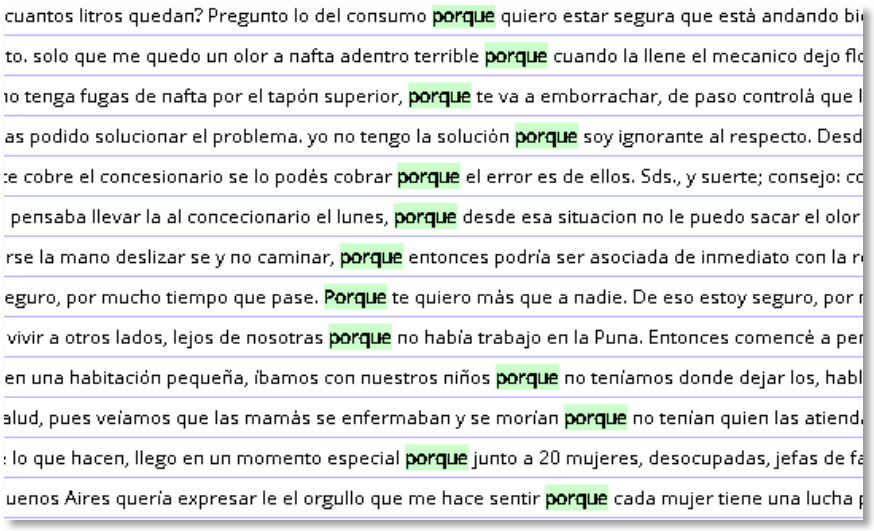
El resto de los casos que siendo identificados por el programa no queden cubiertos con estos patrones activarán un aviso lingüístico que aportará información sobre los usos, contextos y significados de la forma que el programa

identifique. El contenido de estos avisos se expondrá en el último epígrafe de este bloque.

16.4.3.2 *Porque*

La forma *porque* es, categorialmente, una conjunción causal siempre átona que se emplea para encabezar las respuestas que se ofrecen a las preguntas introducidas por *por qué.* Semánticamente es equivalente otras conjunciones cuyo sentido es *puesto que, dado que, ya que* o *pues; Llegó exhausto porque* {*pues/puesto que/ya que/dado que*} *nadie le ofreció ayuda por el camino*.

La multiplicidad de contextos y cotextos que pueden acoger a estas formas hace que sea imposible —con las herramientas que están a nuestra disposición— diseñar patrones que inequívocamente le den un tratamiento correcto a esta forma. Una muestra extraída del Corpus revela la dificultad —cuando no imposibilidad— de corrección certera de estas formas.


cuantos litros quedan? Pregunto lo del consumo **porque** quiero estar segura que está andando bi
to. solo que me quedo un olor a nafta adentro terrible **porque** cuando la llene el mecanico dejo flo
ıo tenga fugas de nafta por el tapón superior, **porque** te va a emborrachar, de paso controlá que l
as podido solucionar el problema. yo no tengo la solución **porque** soy ignorante al respecto. Desd
:e cobre el concesionario se lo podés cobrar **porque** el error es de ellos. Sds., y suerte; consejo: co
pensaba llevar la al concecionario el lunes, **porque** desde esa situacion no le puedo sacar el olor
rse la mano deslizar se y no caminar, **porque** entonces podría ser asociada de inmediato con la r
eguro, por mucho tiempo que pase. **Porque** te quiero más que a nadie. De eso estoy seguro, por
vivir a otros lados, lejos de nosotras **porque** no había trabajo en la Puna. Entonces comencé a per
en una habitación pequeña, íbamos con nuestros niños **porque** no teníamos donde dejar los, habl
alud, pues veíamos que las mamás se enfermaban y se morían **porque** no tenían quien las atiend
: lo que hacen, llego en un momento especial **porque** junto a 20 mujeres, desocupadas, jefas de fa
uenos Aires quería expresar le el orgullo que me hace sentir **porque** cada mujer tiene una lucha

Como puede observarse, tanto la posición de la forma —inicio de frase, seguida de coma, intercalada en el texto, etc.—, como el cotexto izquierdo —infinitivo, forma verbal flexionada, sustantivos, pronombres, adjetivos, etc.— presentan un campo tan amplio y profuso que impide la correcta identificación de estas formas cuando se desarrollan en el eje sintagmático.

Ante esta perspectiva, el programa se limitará a ofrecer un aviso lingüístico que aporte información sobre el uso, contexto y ortografía de esta forma.

16.4.3.3 *Porqué*

Porqué es un sustantivo masculino con flexión de plural cuyo significado es equivalente a *causa*, *motivo* o *razón*. Como sustantivo que es, sus dominios típicos serán los propios de cualquier otro sustantivo, esto es, los de núcleo de un sintagma nominal precedido por uno o más determinantes; *Todos tus porqués carecen de sentido.*

Los patrones que con certeza dan un tratamiento correcto a secuencias que deben involucrar esta forma sustantiva y presentan otras alternativas diferentes son los siguientes:

`@{det}@@{masc}+ @{adj}? @porque`	`[-porque][+porqué]`
**No me vale tu porque.*	

`@{det}@@{masc}+ @{adj}? por que`	`[-por que][+porqué]`
**No me vale tu por que.*	

`@{det}@@{masc}+ @{adj}? por qué`	`[-por qué][+porqué]`
**No me vale tu por qué.*	

`@{det}@@{masc}+ @{adj}? por qués`	`[-por qués][+porqués]`
**No me valen tus por qués.*	

`@{det}@@{masc}+ @{adj}? por ques`	`[-por ques][+porqués]`
**No me valen tus por ques.*	

`@{det}@@{masc}+ @{adj}? @porques`	`[-porques][+porqués]`
**No me valen tus porques.*	

`del @{det}@@{masc}? @{adj}? por qué`	`[-por qué][+porqué]`
**Dame una explicación del por qué de tu actitud.*	

`del @{det}@@{masc}? @{adj}? por que`	`[-por que][+porqué]`
**Dame una explicación del por que de tu actitud.*	

`del @{det}@@{masc}? @{adj}? @porque`	`[-porque][+porqué]`
**Dame una explicación del porque de tu actitud.*	

Como en los casos abordados de la forma segmentada *por qué*, ante la identificación por parte del programa de alguna de estas secuencias además del tratamiento sobre el texto —sustitución de la forma errónea por la correcta—, el programa ofrecerá al usuario la siguiente glosa:

> Cuando queremos hacer referencia al sustantivo *porqué*, que es sinónimo de *causa, motivo* o *razón*, la forma correcta es la que tiene la grafía compacta y acentuada. Se diferencia de las otras formas porque suele ir precedida de un determinante, tiene flexión en plural y puede ser sustituida por la causa, el motivo o la razón: *No sabes el porqué de sus lamentos.*

16.4.3.4 *El aviso lingüístico*

Las secuencias compuestas por *por* y *que* —y sus posibles variantes formales— que el programa detecte en el texto y no se identifiquen con ninguno de los patrones de error codificados serán objeto de un aviso lingüístico.

porque	*Porque* es una conjunción causal átona (nunca lleva tilde), que encabeza las respuestas a preguntas, tanto directas como indirectas, encabezadas por *por qué*. En estos casos, esta forma es sustituible por otras estructuras como ya que, pues, puesto que: Volvió a casa porque/ya que necesitaba saber la verdad. Otro uso de *porque*, además de este causal, es el de valor final: *El anfitrión se esforzó porque todos estuviéramos confortables*. En estos casos, la sustitución puede ser por *para que*. Aunque la Academia acepta tanto la grafía junta como la segmentada, recomienda la primera.
por qué	*Por qué* es una estructura compuesta por la preposición por y el interrogativo tónico *qué*. La utilizamos para introducir preguntas y exclamaciones, tanto directas: *¿Por qué te vas?*, como indirectas: *Me preguntó por qué te ibas*, la forma correcta es la grafía segmentada compuesta por la preposición por y el interrogativo tónico *qué*. En el caso de las interrogativas, la estructura *por qué* admite que añadamos la palabra *razón*, que es inadmisible siguiendo a la grafía compacta: *¿Por qué causa te fuiste?*

por que	*Por que* tiene dos valores: por un lado, la preposición *por* y el pronombre relativo *que*, que se reconoce porque es posible anteponer el artículo al relativo: *No conocía las razones por (las) que se fue de casa*. Por otro lado, esta construcción puede interpretarse como la unión de la preposición y una conjunción subordinante: *Se esmeró por que todo estuviera perfecto*. Esta construcción aparece con verbos, sustantivos y adjetivos que necesitan un complemento introducido por la preposición *por* y llevan, además, una subordinada introducida por la conjunción *que*. Para identificar estos casos, se sustituye la secuencia que sigue a *por*, por el pronombre *eso*: *Tenía preocupación por [que todo el mundo le descubriese]/[eso]*.
@porqué	*Porqué* hace referencia al sustantivo *porqué*, sinónimo de *causa, motivo, razón.* La forma correcta es con la grafía compacta y acentuada. Se diferencia de las otras formas porque suele ir precedida de un determinante, tiene flexión en plural y puede ser sustituida por *la causa, el motivo, la razón*: *No sabes el porqué de sus lamentos*.

Como se puede advertir, se ha diseñado un aviso lingüístico complementario para la forma intratada *por que,* entendida esta como una preposición seguida, bien de un pronombre de relativo; *Desconozco el caso por* (el) *que se le imputa,* bien una conjunción subordinante; Se *alegra por que no sabe lo que le espera*. Como en otros casos, la pluralidad de sentidos y contextos que este compuesto puede ocupar hace imposible llevar a cabo un tratamiento certero que vaya más allá del aviso lingüístico.

16.4.4 *Sino*; *si no*

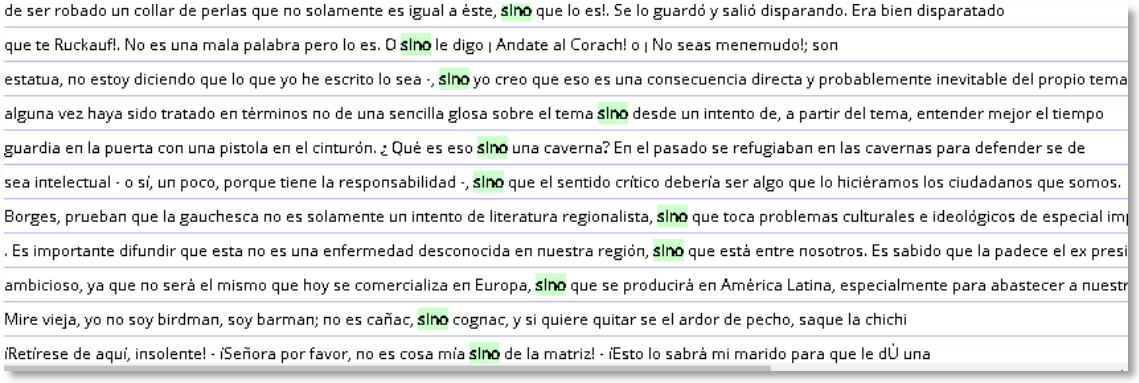
Otro de los casos para los que nuestra lengua cuenta con dos posibles formas que dan lugar a error o, cuando menos, a titubeos gráficos y dudas es el par *sino-si no*. Una, la versión compacta, es ambigua en cuanto a que asume dos categorías y funciones diferentes; la de sustantivo; *El sino de Don Álvaro estaba escrito desde la primera página*; y, por último, la de conjunción adversativa; *No vino sino a incordiar.*

Por otro lado, la forma segmentada surge de la unión de la conjunción *si* con la negación *no.* Esta expresión compuesta que se utiliza para introducir una condición negativa; *No llegarás a ningún lado si no cambies esa actitud*, es la que suele causar vacilaciones ortográficas con la conjunción adversativa; *Con esa actitud no llegarás sino a la cárcel.*

Esta última equivale, semánticamente, a conjunciones como (i) pero sí, también; Lo malo no es lo que dice, sino cómo lo dice; o (ii) excepto; más que; No llamó sino para pedir dinero.

Por su lado, la forma compuesta de conjunción apoyada en una negación debe utilizarse en el sentido de *si acaso no; Si no llegas pronto, te quedarás en tierra.*

Los contextos que pueden asumir estas tres formas diferentes en cuanto a categoría, semántica y ortografía alcanzan, prácticamente, todos los territorios sintácticos. Una muestra extraída del Corpus ilustra la amplitud de contextos que estas formas pueden presentar:


de ser robado un collar de perlas que no solamente es igual a èste, sino que lo es!. Se lo guardó y salió disparando. Era bien disparatado
que te Ruckauf!. No es una mala palabra pero lo es. O sino le digo ¡ Andate al Corach! o ¡ No seas menemudo!; son
estatua, no estoy diciendo que lo que yo he escrito lo sea -, sino yo creo que eso es una consecuencia directa y probablemente inevitable del propio tema
alguna vez haya sido tratado en tèrminos no de una sencilla glosa sobre el tema sino desde un intento de, a partir del tema, entender mejor el tiempo
guardia en la puerta con una pistola en el cinturón. ¿ Què es eso sino una caverna? En el pasado se refugiaban en las cavernas para defender se de
sea intelectual - o sí, un poco, porque tiene la responsabilidad -, sino que el sentido crítico debería ser algo que lo hiciéramos los ciudadanos que somos.
Borges, prueban que la gauchesca no es solamente un intento de literatura regionalista, sino que toca problemas culturales e ideológicos de especial im
. Es importante difundir que esta no es una enfermedad desconocida en nuestra región, sino que está entre nosotros. Es sabido que la padece el ex presi
ambicioso, ya que no será el mismo que hoy se comercializa en Europa, sino que se producirá en América Latina, especialmente para abastecer a nuestr
Mire vieja, yo no soy birdman, soy barman; no es cañac, sino cognac, y si quiere quitar se el ardor de pecho, saque la chichi
¡Retírese de aquí, insolente! - ¡Señora por favor, no es cosa mía sino de la matriz! - ¡Esto lo sabrá mi marido para que le dÙ una

Aunque la intuición lingüística ofrece la posibilidad de poder hacer un marcaje de estas formas mediante otros términos que a simple vista las acoten, la realidad lingüística se impone. Sirva como ejemplo *también*, que en principio parece coaparecer exclusivamente con la forma compacta en función de conjunción adversativa:

Se concluyó con la necesidad no solo de educar a la población sino ***también*** *de proporcionarle recursos.*

Llámame por la tarde, si no ***también*** *puedes intentarlo mañana.*

Pobre Inés, su sino ***también*** *fue trágico.*

Ante este panorama, y junto con las limitaciones de interpretación semántica de la máquina, se ha llevado a cabo un estudio exhaustivo de los cotextos, cuyo resultado ha sido infructuoso, que se ha materializado en el desarrollo de cinco

patrones de error y tres avisos lingüísticos que ofrecen la información necesaria para que el escritor asuma la tarea de corrección.

Para los patrones de error se ha recurrido a valores como la posición de la secuencia en la frase y otros elementos ortotipográficos como las comas y las mayúsculas, que sirven para ceñir las formas y asegurar de un modo certero la intención comunicativa del escritor bien tenga intención de establecer una relación adversativa entre dos partes o por el contrario busca la expresión de una condición negativa.

Se exponen, a continuación, los patrones codificados tras comprobar que estos contextos son privativos y excluyentes de cada variante:

no solo \w+, si no	[-si no][+sino]
**No solo me resulta aburrido si no desagradable.*	

no \w+ si no	[-si no][+sino]
**No quería trabajar ni estudiar si no vivir eternamente de sus padres.*	

ni \w+ si no	[-si no][+sino]
**No quería trabajar ni estudiar si no vivir eternamente de sus padres.*	

@{det}@@{masc} @{adj}? si no	[-si no][+sino]
**Hacer y deshacer entuertos, ese era su fatal si no.*	

sino #{vb}	[-sino][+si no]
**Sino dimite ya, deberían cesarlo.*	

En cuanto a los avisos lingüísticos, se han diseñado las siguientes entradas para dar asistencia a aquellas secuencias protagonizadas por estas formas que no han sido tratadas por los patrones precedentes:

no \w+ si no	Cuando *sino* introduce un elemento que reemplaza al negado en la oración precedente, la grafía correcta es la compacta: *No vi toda la exposición sino lo más importante*. El valor de esta conjunción adversativa puede variar según el enunciado; además de contraponer una idea afirmativa a

no \w+ sino	otra negativa expresada anteriormente, en algunos casos denota adición enfática, mientras que en otros puede sustituirse por *más que, salvo, excepto,* etc. No debe confundirse esta conjunción adversativa *sino*, con la secuencia de conjunción condicional y negación *si no*, donde la negación es tónica, a diferencia del [*no*] átono de *sino*: *No sé si no tendrá que dimitir*.
sinó	No debe confundirse la conjunción adversativa *sino*, con la secuencia de conjunción condicional y negación *si no*, donde la negación es tónica, a diferencia del [*no*] átono de *sino*: *No sé si no tendrá que dimitir*.

16.5Vulgarismos

Aunque es evidente que la intencionalidad que supone la escritura de un texto ya exige cierta pericia y ahuyenta los vulgarismos que forman parte del repertorio de cualquier usuario, se han registrado unos cuantos patrones que, por su recurrencia, justifican su inclusión en nuestro repertorio. Formas erróneas como **inclusives*, **el taxis*, **la apéndiz,* **usté*, y expresiones como **Hizo dos llamadas de la que vino*; **Contra más se esfuerza, peores son sus resultados*; **Te se nota poco la dieta,* serán tratadas en PatErr.

De esta diversidad de afección surge la variedad de fenómenos que se le pueden asociar a estos patrones; algunos presentan inconsistencias en la construcción de una expresión, otras atañen a errores morfológicos, unos se consideran arcaísmos y otros simplemente vulgarismos que deben evitarse en todo caso. De aquí se colige que la detección e intervención sobre estos términos puede incidir sobre varios niveles de la lengua; morfológico, ortográfico, sintáctico y, en último caso, estilístico.

Cabe destacar, dentro de este universo de voces censuradas por la norma culta, un nutrido grupo de verbos y formas verbales que por su irregularidad o alternancia vocálica generan formas incorrectas; ⊗*andé,* ⊗*fregas,* ⊗*callarsen* o ⊗*esperaros*. Estas formas también han recibido tratamiento correctivo a partir de un patrón de error.

A continuación, se ofrece una muestra de algunos de los patrones codificados que reciben esta consideración.

usté	usted
se te	se te
se me	se me
inclusives	inclusive
contra más	mientras más cuanto más
#barajear	[-baraje][+baraj]
@{det}@@{masc sg} @{adj}? taxis	[-taxis][+taxi]
{prep} taxis	[-taxis][+taxi]
de que @{vb},	[-de que @{vb}] [+al {inf}]| [+una vez que @{vb},]| [+cuando @{vb},]| [+en cuanto @{vb},]
a lo que @{vb} a la que @{vb}	[-a lo que] [+ una vez que @{vb},]| [+ en cuanto @{vb},]| [+ cuando @{vb},]

Como se puede intuir, dentro de este repertorio hay patrones de error que se corregirán automáticamente y otros de recomendación que, como en el caso de los dos últimos, proporcionarán al usuario varias opciones adecuadas para que él de forma interactiva pueda subsanar su texto.

Conclusiones y horizontes futuros

Conclusiones

PLANTEAMIENTO DE UN SISTEMA VIABLE Y COMPETITIVO PARA LA REVISIÓN TEXTUAL DEL ESPAÑOL BASADO EN TÉCNICAS DE BAJO NIVEL

Las conclusiones que pueden extraerse de todo el trabajo aquí propuesto pueden concretarse en la constatación de que a partir de la reutilización de recursos propios de las técnicas de bajo nivel del PLN como un lematizador, un flexionador, un conjugador y un lexicón que ofrezca garantías en términos de cobertura y rigor en la etiquetación, se puede desarrollar un *sistema viable de revisión textual automática* que, basado en la identificación de patrones de error en el texto susceptible de revisión, y prescindiendo de onerosas técnicas de alto nivel de abstracción y análisis en términos lingüísticos, puede resultar útil y competitivo en el entorno de la revisión textual automática[104].

CREACIÓN DE UN LENGUAJE DE CODIFICACIÓN EXPRESIVO Y GENERATIVO QUE POSIBILITA LA TRANSFERENCIA DE CONOCIMIENTOS LINGÜÍSTICOS AL ENTORNO COMPUTACIONAL

Por otro lado, se ha presentado un *modelo de codificación de patrones* de error con un lenguaje expresivo y generativo que ha propiciado la comunicación entre el lingüista, y sus abstracciones fruto de la investigación empírica, y la máquina y sus limitaciones. Esta herramienta, permite crear patrones sincréticos de gran alcance válidos, por un lado, para llevar a cabo la generación automática de toda la casuística posible derivada de un mismo fenómeno, y por otro, para crear estructuras globales figurativas que puedan identificarse en un texto etiquetado. Con independencia de cuál sea el enfoque que se adopte para la explotación de los patrones, se ha conseguido crear un sistema globalizador de la casuística de errores consistente en términos de cobertura.

[104] El prototipo al que aludimos en la Introducción así lo demuestra.

UNA METODOLOGÍA DE TRABAJO CON UN ENFOQUE PROPIO: LA LINGÜÍSTICA DE ERRORES CON FINES COMPUTACIONALES

Junto con estos recursos, se ha propuesto una metodología de trabajo inserta en lo que hemos denominado *lingüística de errores con fines computacionales* que pretende, desde el enfoque del error, ofrecer bases sólidas —tanto teóricas como prácticas— para las tareas que debe acometer el lingüista en cuanto a la identificación, circunscripción, análisis y formalización de los errores del español que se registran insertos en el discurso escrito.

DESARROLLO DEL RECURSO. UN COMPLEMENTO PARA LOS PROGRAMAS DE VERIFICACIÓN TEXTUAL EN ESPAÑOL

A lo largo de estas páginas se ha ido acotando el alcance real de PatErr, sus limitaciones y posibilidades futuras, bien en el entorno comercial, bien en el ámbito didáctico. En el estado de desarrollo en el que se presenta, una de sus derivaciones puede concretarse en servir como *complemento para otros programas correctores* que presentan intervención en los niveles ortográficos y gramaticales —concordancias—. Los análisis contrastivos que se han llevado a cabo con otros recursos de corrección automática para el español insertos en el mercado muestran la potencialidad en términos de cobertura y prestaciones —servicio de asesoría— que la propuesta que aquí presentamos puede ofrecer.

Este punto de madurez incipiente del recurso en términos cuantitativos, no obsta para poder pensar en el futuro en un *programa independiente* capaz de asumir en solitario estas tareas, valentía infundada por la viabilidad de los planteamientos, la potencialidad del propio recurso y la robustez del sistema en el sentido de no colapsar ante secuencias no registradas o previstas en el sistema y con una capacidad constante de revisión, ampliación y actualización de la información.

Trabajos futuros

A lo largo de esta exposición se han ido abriendo líneas de investigación nuevas con retos de diversa magnitud que culminarían con el desarrollo de un recurso independiente de amplia cobertura para la revisión textual del español. Estas líneas de investigación y desarrollo, en el ámbito de la ingeniería lingüística, pueden sintetizarse en los siguientes hitos:

UN SISTEMA DE DESAMBIGUACIÓN

El desarrollo de un sistema capaz de minimizar, e idealmente eliminar, las ambigüedades en el plano gramatical —categorías— garantizaría que la aplicación de los patrones sobre un texto opera sobre las unidades adecuadas. A pesar del recurso auxiliar —en términos de desambiguación— que ha supuesto la inclusión de cotexto en los patrones de error, son varios los casos registrados en los que, estando ceñido el error, la presencia de la ambigüedad dificulta la aplicación precisa de los patrones. Recuérdense casos como *la venda; la menta; la compra,* etc., que aceptan dos configuraciones sintácticas diferentes.

Este sistema desambiguador habría de operar una vez generados los patrones de error apartando del repertorio de cada fenómeno tratado aquellos casos que presentan la posibilidad de adoptar más de una configuración sintáctica en términos categoriales; «det + N»; «pn + N». Excluidos estos, se sacrificará la cobertura del sistema —estos casos no serán tratados—, pero se evitará la corrección sobre secuencias correctas que quedan fuera del ámbito de aplicación del patrón. Un desambiguador permitiría, por un lado, acotar el radio de acción de los patrones y, por otro lado, serviría para depurar los resultados derivados del sistema generativo que se presenta. Un enfoque que consideramos productivo para la consecución de este programa es el desarrollado a partir de N-gramas, que ha ofrecido resultados satisfactorios en otros proyectos inscritos en el PLN.

ESTUDIO DE APLICABILIDAD DE PATRONES. DEPURACIÓN DE LOS RESULTADOS

Con relación a la generación masiva de patrones, sobre todo en aquellos casos en los que el producto de la generación son estructuras que presentan complementos, será conveniente hacer un estudio minucioso de la aplicabilidad de cada patrón y, por otro lado, de su utilidad y frecuencia de uso. Como ha quedado expuesto, las producciones derivadas del sistema generativo que aquí proponemos copan toda la casuística de muchos de los fenómenos tratados mediante la generación irrestricta de errores derivados. Muchos de estos presentarán inconsistencias cuando menos, en términos semánticos; **Pensé casi de que tornillos*, por lo que puede resultar de gran utilidad emprender un estudio empírico sobre un corpus con el fin de depurar la masa de errores generados y extraer aquellos más frecuentes y probables. Un recurso que puede servir para estos fines depurativos puede ser, una vez más, la técnica de N-gramas.

MÓDULO AUTÓNOMO DE CORRECCIÓN ORTOGRÁFICA

Otro recurso que en términos cualitativos permitiría el despegue de PatErr como recurso integrado en las industrias de la lengua sería el desarrollo de un sistema de corrección ortográfica eficaz que garantice una primera revisión y asistencia sobre el texto con el fin de subsanar los errores que se constatan en este plano.

A partir de este, la aplicación de los patrones de error que se han ofrecido actuará con todas las garantías y certidumbre sobre palabras y estructuras que hayan pasado ese filtro ortográfico. El desarrollo de este corrector, no obstante, no evita la concurrencia de los patrones de ortografía que aquí se han expuesto; como se ha señalado, una amplia mayoría de los correctores ortográficos integrados en las industrias de la lengua son incapaces de detectar los errores generados en el texto escrito cuando el término que genera la incidencia es un homófono o presenta diferentes opciones en relación con la segmentación.

CREACIÓN DE UN MÓDULO QUE IDENTIFIQUE DISCONCORDANCIAS

La asistencia de un sistema capaz de detectar los desvíos relacionados con las concordancias, tanto intersintagmáticas como intrasintagmáticas, complementaría la acción de PatErr y permitiría abordar casos que, en el punto de desarrollo actual, quedan fuera de su alcance. No obstante, como en el caso de la corrección ortográfica, este desarrollo no implica la postergación de los patrones que se han ofrecido con relación a los fracasos en la concordancia, que como se ha constatado, dan cobertura a fenómenos que pasan desapercibidos para otros revisores textuales. Los patrones de PatErr, que contemplan casos particulares en los que suelen concurrir estos errores, servirán de complemento a ese motor principal capaz de controlar estas relaciones que establecen las unidades de la lengua en el discurso.

El desarrollo de estos dos sistemas, con independencia del enfoque que se asuma para su consecución —patrones de error, reglas lingüísticas legítimas, datos, técnicas estadísticas, etc.— actuarían en una fase previa a la revisión de patrones para *normalizar* el texto y ofrecerlo sin interferencias que anulen la validez de los patrones que se han diseñado.

Por último, como puede sospecharse, el trabajo inmediato que surge en el día a día de este proyecto es el diseño y desarrollo manual de más patrones que enriquezcan y fortalezcan este recurso con los errores que con más frecuencia salpican nuestros textos.

Referencias

ABNEY, S. P. (1987). *The English Noun Phrase in its Sentential Aspect*. Cambridge: Cambridge Mass.

ALCÁNTARA PLÁ, M. (2007). *Introducción al análisis de estructuras lingüísticas en corpus: aproximación semántica*. Madrid: Servicio de Publicaciones de la Universidad Autónoma de Madrid.

ALCARAZ VARÓ, E. (1990). *Tres paradigmas de la investigación lingüística*. Alcoy: Marfil.

ALMEIDA, M. (2007). Tres tesis sobre el dequeísmo. *Revista De Filología.* 49-57.

ALONSO, A.; HENRÍQUEZ UREÑA, P. (1999). *Gramática castellana* (30 ª ed.). Buenos Aires: Losada.

ALVAR EZQUERRA, M. (1994). *Diccionario de voces de uso actual*. Madrid: Arco Libros.

BAKER, P.; HARDIE, A. Y MC ENERY, T. (2006). *A glossary of corpus linguistics.* Edinburgh: University Press.

BALARI, S. (1999). Formalismos gramaticales de unificación y procesamiento basados en restricciones *Revista española de lingüística aplicada*. (1), 117-152.

BIBER, D.; CONRAD, S.; REPPEN, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

BOSQUE, I. (1979). Perspectivas de una lingüística no discreta. En Abad, F. *et al.* (eds.) *Metodología y gramática generativa*. (pp. 81-110) Madrid: SGEL.

BOSQUE, I. (1987). Construcciones morfológicas sobre la coordinación. *Lingüística Española Actual.* 9(1), 83-100.

BOSQUE, I. (1989). *Las categorías gramaticales; relaciones y diferencias.* Madrid: Síntesis.

BOSQUE, I. Y DEMONTE, V. (1999). *Gramática descriptiva de la lengua española*. Madrid: Real Academia Española, Espasa Calpe.

BOSQUE, I. Y GUTIÉRREZ-REXACH, J. (2009). *Fundamentos de sintaxis formal*. Madrid: Akal.

BRIZ GÓMEZ, A. (1989). *Sustantivación y lexicalización en español: (la incidencia del artículo)*. Valencia: Universidad de Valencia.

CALZOLARI, N.Y PICCHI, E. (1994). A lexical workstation: from textual data to structured database. En Atkins, B.T.S y Zampolli, A. (eds.). *Computational approaches to the lexicon.* (pp. 439-467). Oxford: Oxford University Press.

CARRERAS RIUDAVETS, F. J.; RODRÍGUEZ RODRÍGUEZ, G.; HERNÁNDEZ FIGUEROA, Z. (2011). Conjugador TIP de verbos del español. En *Actas del XII Simposio Internacional de Comunicación Social*. Vol I. Santiago de Cuba.

CARRERAS RIUDAVETS, F.; RODRÍGUEZ DEL PINO, J.; HERNÁNDEZ FIGUEROA, Z.; RODRÍGUEZ RODRÍGUEZ, G. (2012). A morphological analyzer using hash tables in main memory (MAHT) and a lexical knowledge base. *Computational Linguistics and Intelligent Text Processing.* (80-91). Berlín-Heidelberg: Springer.

CASTRO CASTRO D. (2012). *Métodos para la corrección ortográfica automática del español.* (Tesis de Máster en Ciencia de la Computación). Universidad de Oriente, Santiago de Cuba.

CERVERA RODRÍGUEZ, A. (2011). Teoría lingüística actual en la NGLE. *Revista Cálamo FASPE.* (57) 14-21.

CHACÓN BELTRÁN, R. C. (2008). El uso de expresiones regulares en la detección de errores escritos: Implicaciones para el diseño de un corrector gramatical. En *Actas Del VIII Congreso De Lingüística General; El Valor De La Diversidad (Meta) Lingüística.*

CHOMSKY, N. (1982). *Rules and representations*. Oxford: Basil Blackwell.

CHOMSKY, N. (1995). *The minimalist program.* Cambridge: Cambridge University Press.

CIVIT TORRUELLA, M. (2003). *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*. Colección de monografías, 3. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).

CLAVE, SM. (1997). *Diccionario de Uso del Español Actual*. Consultado en http://clave.smdiccionarios.com/app.php. [último acceso 15-06-2015].

DEMONTE, V. (1991). *Detrás de la palabra.* Madrid: Alianza,

DUCROT, O. Y TODOROV, T. (1974). *Diccionario enciclopédico de las ciencias del lenguaje: Lingüística.* (Trad. E. Pezzoni). Buenos Aires: Siglo XXI.

EGUREN GUTIÉRREZ, L. (1989). Algunos datos del español en favor de la hipótesis de la frase determinante. *Revista Argentina De Lingüística.* 5, (1-2), (163-203).

EGUREN GUTIÉRREZ, L. Y FERNÁNDEZ SORIANO, O. (2006). *La terminología gramatical*. Madrid: Gredos.

ESCANDELL VIDAL, V. (2004). *Fundamentos de semántica composicional.* Barcelona: Ariel.

ESCANDELL VIDAL, V. Y LEONETTI, M. (2000). Categorías funcionales y semántica procedimental. En *Actas del V Congreso de Lingüística General*. (1727-1738). Madrid: Arco.

ESCARPANTER, J. (1994). *Eso no se escribe así: Los 1000 errores más frecuentes en español*. Madrid: Playor.

FERREIRA, A.Y KOTZ, G. (2010). ELE-Tutor Inteligente: Un analizador computacional para el tratamiento de errores gramaticales. Español como Lengua Extranjera. *Revista Signos.* 43(73), 211-236.

FRIEDL, J. (2006). *Mastering Regular Expressions*, Estados Unidos: O'Reilly.

GALLÉS, n. s. (2000). *LEXESP: Léxico informatizado del español.* Barcelona: Edicions Universitat Barcelona.

GARRIDO, J. (1996). Sintagmas nominales escuetos. En Bosque, I. (ed.) *El sustantivo sin determinación. La ausencia de determinante en la lengua española.* (pp. 269-338). Madrid: Visor Libros.

GIVON, T. (1989). *Mind, code and context: Essays in pragmatics*. Sussex: Erlbaum, Lawrence.

GÓMEZ GUINOVART, X. (1996). Aportaciones a la metodología de evaluación de los sistemas de verificación automática de la sintaxis. *Procesamiento Del Lenguaje Natural,* (19), 7-13.

GÓMEZ GUINOVART, X. (2000). Lingüística computacional. En Ramallo *et al.* (Eds.),*Manual de ciencias da linguaxe.* (pp. 221-268). Barcelona: UOC Universitat Oberta de Catalunya.

GÓMEZ GUINOVART, X. (2001). Representación y procesamiento de la gramática léxico-funcional. *Novatica.* Secciones Técnicas. (150), 57.

GÓMEZ GUINOVART, X.; MASALLES CASTELLÓN, I. Y OLIVER GONZÁLEZ, A. (2010). *Lingüística computacional.* Barcelona: UOC Universitat Oberta de Catalunya.

GRISHMAN, R. (1986). *Introducción a la lingüística computacional*. (Trad. Moreno Sandoval, A. 1991). Madrid: Visor.

GUTIÉRREZ ORDÓÑEZ, S. (1994). Problemas en torno a las categorías funcionales. En Hernández Paricio, F. (ed.) *Perspectivas sobre la Oración.* (pp. 71-99). Zaragoza: Grammaticalia.

HARRIS, Z. S. (1946). From morpheme to utterance. *Language.* 22(3), 161-183.

HOCKETT, C. F. (1958). *Curso de Lingüística moderna*. (Trad. de 1971). Buenos Aires: Eudeba.

IDE, N.; LE MAITRE, J. Y VÉRONIS, J. (1993). Outline of a model for lexical databases. *Information Processing & Management.* 29(2), 159-186.

LAROUSSE. (1997). *Gran Diccionario de la Lengua Española Larousse.* Barcelona: Planeta.

LÁZARO CARRETER, F. (1980). *Estudios de lingüística.* Barcelona: Crítica.

LEECH, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing.* 8 (4), 275-281.

LEECH, G. (2005). Adding linguistic annotation. En Wynne, M. (ed.). *Developing Linguistic Corpora: A Guide to Good Practice*. (pp. 17-29). Oxford: Oxbow Books.

LEECH, G. Y WILSON, A. (1996). EAGLES. Recommendations for the syntactic annotation of corpora. Consultado en http://www.ilc.cnr.it/EAGLES/segsasg1/. [Acceso 25-05-2017].

LEONETTI JUNGL, M. (1990). *El artículo y la referencia*. Madrid: Taurus.

LEONETTI JUNGL, M. (2000). *Los determinantes*. Madrid: Arco Libros.

LIDDY, E. D. (2001). Natural language processing. En Marcel Decker, I. (Ed.), *Encyclopedia of library and information science* (2ª ed.). New York: Marcel Decker, Inc.

LLISTERRI, J. (2003). Lingüística y tecnologías del lenguaje. *Lynx. Panorámica De Estudios Lingüísticos.* (2) 9-71.

LLISTERRI, J. (2007). El español y las nuevas tecnologías. En *Lingüística Aplicada del español.* (pp. 483-520). Madrid: Arco Libros.

LLISTERRI, J.; MARTÍ, M. A. (2004) *Tecnologías del texto y del habla.* Barcelona: Publicacions i Edicions de la Universitat de Barcelona.

MAIRAL USÓN, R. Y GONZÁLVEZ-GARCÍA, F. (2010). Verbos y construcciones en el espacio cognitivo-funcional del siglo XXI. En *Val Álvaro y Horno Chéliz: La gramática del sentido: léxico y sintaxis en la encrucijada.* Zaragoza: Universidad de Zaragoza.

MAIRAL USÓN, R.; PEÑA CERVEL M. S.Y RUÍZ DE MENDOZA F. J. (2012). *Teoría lingüística: métodos, herramientas y paradigmas.* Madrid: Editorial Universitaria Ramón Areces.

MARCOS MARÍN, F. (1994). *Informática y humanidades*. Madrid: Gredos.

MARCOS MARÍN, F.; SATORRE GRAU, F.J. Y VIEJO SÁNCHEZ, M. L. (1998). *Gramática española*. Madrid: Síntesis.

MARTÍ ANTONÍN, M. A. (1999). Panorama de la lingüística computacional en europa. *Revista Española De Lingüística Aplicada,* (1), 11-24.

MARTÍ ANTONÍN, M. A.; LLISTERRI, J. (2002). *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita*. Barcelona: Ediciones Universitat de Barcelona.

MARTÍ, M. A.Y TAULÉ, M. (2011). La Academia y la investigación universitaria en las tecnologías de la lengua. En Senz y Alberte (Eds.), *El dardo en la academia. esencia y vigencia de las academias de la lengua española.* (pp. 1195-1242).

MARTÍN DE SANTA OLALLA SÁNCHEZ, A. (1999). Propuesta de un modelo de codificación morfosintáctica para corpus de referencia en lengua española. *Estudios De Lingüística Del Español, ELiEs.* (3). (Microforma).

MARTÍN ZORRAQUINO, M. A. Y PORTOLÉS, J. (1999) Los marcadores del discurso. En Bosque, I y Demonte, V. (eds): *Gramática descriptiva de la lengua española.* (pp. 4051 – 4212). Madrid: Real Academia Española, Espasa Calpe.

MARTÍNEZ DE SOUSA, J. (1992). *Dudas y errores de lenguaje.* Madrid: Editorial Paraninfo.

MARTÍNEZ DE SOUSA, J. (2000). *Manual de estilo de la lengua española*. Oviedo: Trea.

MARTÍNEZ DE SOUSA, J. (2004). *Ortografía y ortotipografía del español actual*. Oviedo: Trea.

MC ENERY, T. (2003): Corpus Linguistics. En Mitkov, R. (ed.). *The Oxford Handbook of Computational Linguistics.* (pp. 448-463) Oxford: Oxford University Press.

MC ENERY, T.; WILSON, A. (1996). *Corpus linguistics*. Edimburgo: Edinburg University Press.

MC ENERY, T.; WILSON, A. (2001). *Corpus linguistics: an introduction*. Edimburgo: Edinburgh University Press.

MOLINER, M. (1996). *Diccionario de Uso del Español.* Madrid: Gredos.

MONACHINI, M., & CALZOLARI, N. (1996). *EAGLES synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages*. París: Centre National de la Recherche Scientifique Paris.

MORENO, L.; PALOMAR, M.; MOLINA, A. Y FERNÁNDEZ, A. (1999). *Introducción al procesamiento del lenguaje natural.* Alicante: Servicio de Publicaciones de la Universidad de Alicante.

MORENO ORTIZ, A. (2000). Diseño e implementación de un lexicón computacional para lexicografía y traducción automática. *Estudios De Lingüística Del Español, 9*.

MORENO SANDOVAL, A. (1998). *Lingüística computacional: Introducción a los modelos simbólicos estadísticos y biológicos*. Madrid: Síntesis.

MORENO SANDOVAL, A. (2001). *Gramáticas de unificación y rasgos*. Madrid: A. Machado Libros.

MOURE, T. Y LLISTERRI, J. (1996) Lenguaje y nuevas tecnologías. El campo de la lingüística computacional. En Fernández Pérez, M. (Coord.) *Avances en lingüística aplicada*. Santiago de Compostela: Universidade de Santiago de Compostela, Servicio de Publicacións e Intercambio Científico. (pp. 147-228).

NABER, D. (2003). *A rule-based style and grammar checker*. (Tesis Doctoral Technische Fakultät, Universität Bielefeld).

NÁÑEZ FERNÁNDEZ, E. (1984). Sobre dequeísmo. *Revista De Filología Románica.* vol. (2); (pp.239-249).

NAZAR, R. Y RENAU, I. (2012). Google books n-gram corpus used as a grammar checker. *Proceedings of the Second Workshop on Computational Linguistics and Writing (CLW 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering.*

OFICINA DE PUBLICACIONES OFICIALES DE LA LAS COMUNIDADES EUROPEAS. (1997). *Lenguaje y tecnología. De la Torre de Babel a la aldea global*. Luxemburgo.

PAREDES GARCÍA, F. (2014). *El libro del español correcto: Claves para hablar y escribir bien en español*. Madrid: Espasa.

PARODI, G. (2008). Lingüística de corpus: Una introducción al ámbito. *RLA, Revista De Lingüística Teórica y Aplicada.* 46 (1), 93.

PARRET, H. (1974). Peter Hartmann. *Discussing Language.* La Haya: Mouton. (151-178).

PAVÓN LUCERO, M. V. (2003). *Sintaxis de las partículas*. Madrid: Visor Libros.

PERIÑÁN PASCUAL, C. (2005) Procesamiento del lenguaje natural: de lingüista a ingeniero del conocimiento. En Brady *et al.* eds.) *Nuevas Tendencias en Lingüística Aplicada*. (pp. 293-317). Murcia: Quaderna.

PERIÑÁN PASCUAL, J. C. (2012). En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica. *Onomázaein* 26 (2), 13-48.

PORTO DAPENA, J. Á. (2004). La lematización de los pronombres en los diccionarios actuales. *Revista de Lexicografía*. (10), 139-182.

QUESADA LÓPEZ, R.; SANTANA PÉREZ, I.; SANTANA SUÁREZ, O.; PÉREZ AGUIAR, J. R. (2008). *Desambiguación funcional según las estructuras sintácticas de la gramática española*. Las Palmas de Gran Canaria: Universidad de Las Palmas de Gran Canaria, Facultad de Informática.

QUIRK, R. (1992). On corpus principles and design. *Svartvik.* 457-469.

RAFEL I FONTANALS, J Y SOLER I BOU, J. (2003). El procesamiento de corpus. En Martí Antonín, M. A. (Ed.) *Tecnologías del lenguaje.* (pp. 41-73). Barcelona: Editorial UOC.

RAMÍREZ BUSTAMANTE, F. Y SÁNCHEZ LEÓN, F. (1996). *GramCheck*: Un corrector gramatical para el español. *Procesamiento Del Lenguaje Natural,* 19, 30.

RAMÍREZ BUSTAMANTE, F.; SÁNCHEZ LEÓN, F. Y DECLERCK, T. (1997). Corrección gramatical y preprocesamiento. *Procesamiento Del Lenguaje Natural,* 21.

RAMÍREZ BUSTAMANTE, F.; SÁNCHEZ LEÓN, F. Y DECLERCK, T. (1998). *CON—TEXT.* Un corrector gramatical de bajo nivel. *Procesamiento Del Lenguaje Natural,* 23.

RAMÍREZ BUSTAMANTE, F.; RODRÍGUEZ SELLÉS, F. Y SÁNCHEZ LEÓN, F. (1994). Tipología de errores gramaticales del español para un sistema automático de corrección de textos. *Actas Del X Congreso De Lenguajes Naturales y Lenguajes Formales.* (pp. 573-580). Sevilla.

REAL ACADEMIA ESPAÑOLA Y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA. (2005). *Diccionario panhispánico de dudas*. Madrid: Santillana.

REAL ACADEMIA ESPAÑOLA Y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA (2009). *Nueva gramática de la lengua española*. Madrid: Espasa Calpe.

REAL ACADEMIA ESPAÑOLA Y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA. (2010). *Nueva gramática de la lengua española: Manual*. Madrid: Espasa Calpe.

REAL ACADEMIA ESPAÑOLA Y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA. (2010). *Ortografía de la lengua española*. Madrid: Espasa Calpe.

REAL ACADEMIA ESPAÑOLA. *Diccionario de la Real Academia Española*. Versión electrónica de la vigésima segunda edición, en proceso de enmienda. Consultado en http://www.rae.es/rae.html. [Acceso 15-06-2012].

ROCA PONS, J. (1960). *Introducción a la gramática.* Barcelona: Vergara Editorial.

RODRÍGUEZ ESPIÑEIRA, M. J.; PENA SEIJAS, J. (2008). *Categorización lingüística y límites intercategoriales.* Santiago de Compostela: Universidad de Santiago.

RODRÍGUEZ RAMALLE, M. T. (2005). *Manual de sintaxis del español*. Madrid: Castalia.

RODRÍGUEZ, H. (2000). Técnicas básicas en el tratamiento informático de la lengua. *Quark, Ciencia, Medicina, Comunicación y Cultura. 19*, (26-34).

ROJO, G. (2002). El empleo de corpus textuales en la investigación lingüística. *Actas Del VIII Simposio de actualización científica y didáctica de lengua española y literatura. 7*(10), 107-127. Sevilla.

ROJO, G. (2002). Sobre la lingüística basada en análisis de corpus [en línea]. *https://www.academia.edu/5946111/Sobre_la_Ling%C3%BC%C3%ADstica_basada_en_el_an%C3%A1lisis_de_corpus.* [último acceso 2/06/2017].

ROJO, G. (2008). Lingüística de corpus y lingüística del español. *Ponencia Plenaria Presentada en El XV Congreso de la Asociación de Lingüística y Filología de América Latina.* Montevideo*.*

RUÍZ ANTÓN, J. C. (2005). Lenguaje e informática. Lenguaje y ordenadores. En López García, A. *et al. Conocimiento y lenguaje.* (pp. 401—453). Valencia: Servicio de Publicaciones de la Universidad de Valencia.

SAN MATEO, A. (2016). Un corpus de bigramas utilizado como corrector ortográfico y gramatical destinado a hablantes nativos de español. *Revista Signos, 49*(90), 94-118.

SÁNCHEZ LEÓN, F. (1995). Desarrollo de un etiquetador morfosintáctico para el español. *Procesamiento Del Lenguaje Natural,* (17), 14-28.

SÁNCHEZ, F.; PORTA, J.; SANCHO, J. L.; NIETO, A.; BALLESTER, A.; FERNÁNDEZ, A. Y RUIZ, R. (1999). La anotación de los corpus CREA y CORDE. *Revista De La Sociedad Española Para El Procesamiento Del Lenguaje Natural, 25*, 175-182.

SANTALLA DEL RÍO, M.P. (2005). La elaboración de corpus lingüísticos. En Cal, Núñez, y Palacios (eds.). *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas.* (pp. 45-66). Santiago: Universidade de Santiago de Compostela.

SANTANA SUÁREZ, O.; CARRERAS RIUDAVETS, F.; FIGUEROA, Z. Y RODRÍGUEZ, G. (2003). The Spanish morphology in internet. En *Web Engineering*. Berlín-Heidelberg: Springer.

SANTANA SUÁREZ O.; CARRERAS RIUDAVETS, F. J.; PÉREZ AGUIAR, J. R. Y RODRÍGUEZ RODRÍGUEZ, G. (2004). Relaciones morfoléxicas prefijales del español. *Procesamiento Del Lenguaje Natural,* 32.

SANTANA SUÁREZ, O.; CARRERAS RIUDAVETS, F. J.; PÉREZ AGUIAR, J. R. (2004). *Relaciones morfoléxicas sufijales para el procesamiento del lenguaje natural*. Madrid: Mileto.

SEBASTIÁN GALLÉS, N. (2000). *LEXESP: Léxico informatizado del español.* Barcelona: Edicions Universitat de Barcelona.

SECO, M. (1999). *Gramática esencial del español: introducción al estudio de la lengua.* Madrid: Espasa-Calpe.

SECO, M.; ANDRÉS, O.; RAMOS, G. (2011). *Diccionario del español actual*. Madrid: Aguilar.

SINCLAIR, J. (1996). Preliminary recommendations on corpus typology. *EAGLES* http://www.ilc.pi.cnr.it/EAGLES/corpustyp/corpustyp.html [último acceso 2/06/2017].

SPERBER, D. Y WILSON, D. (1994). *La relevancia: Comunicación y procesos cognitivos*. Madrid: Visor.

SUBIRATS RÜGGEBERG, C. (2001). *Introducción a la sintaxis léxica del español*. Madrid: Vervuert.

TORRUELLA, J. Y LLISTERRI, J. (1999). Diseño de corpus textuales y orales. *Filología e Informática. Nuevas Tecnologías en los Estudios Filológicos.*

VAL ÁLVARO, J. F. Y MENDÍVIL GIRÓ, J. L. (2011). Concordancia en oraciones escindidas con sujeto pronominal. En Escandell Vidal, V. *et al. 60 Problemas De Gramática dedicados a Ignacio Bosque.* (pp. 299-305). Madrid: Akal.

VAN COMPERNOLLE, R. A. (2009). Terry Nadasdi and Stéphan Sinclair, BonPatron: An online grammar, spelling, and expression checker. Nadaclair Technologies, *Journal of French Language Studies.* 19(03), 406-409.

VARELA ORTEGA, S. (1979). Categorías sintácticas y teoría gramatical. En Abad, F. *et al.* (eds.) *Metodología y gramática generativa*. (pp. 53-80). Madrid: SGEL.

VARELA ORTEGA, S. (1999). Sobre las relaciones de la morfología con la sintaxis. *Revista española de lingüística.* 29(2), 257-281.

VERBERNE, S. (2002). *Context-sensitive spell checking based on word trigram probabilities*. (Tesis de Doctorado inédita). University de Nijmegen.

VILLAYANDRE, M. (2008). Lingüística con corpus. *EH Filología,* 329-349.

VILLENA, J.; GONZÁLEZ, B.; GONZÁLEZ, B. Y MURIEL, M. (2002). *STILUS*: Sistema de revisión lingüística de textos en castellano. *Procesamiento Del Lenguaje Natural.* (29), 305-306.

VOX. (1997). *Diccionario General de la Lengua Española.* Barcelona: Biblograf.

VV. AA. (1979). *Metodología y gramática generativa* [Bosque "Perspectivas de una lingüística no discreta", *Revista española de lingüística*. 7, 1, (1977). 155-177.

VV. AA. (1991). *Gran Diccionario de Sinónimos y Antónimos*. Madrid: Espasa Calpe.

VV. AA. (2013). *Las 500 dudas más frecuentes del español.* Madrid: S.L.U. Espasa.

WEDBJER RAMBELL, O. (1999). *Error Typology for Automatic Proof-Reading Purposes.* (Tesis Doctoral en Lingüística). Uppsala University, Suecia.

WYNNE, M. (2005). *Developing linguistic corpora: A guide to good practice* Oxbow Books.

ZAMPOLLI, A.; CALZOLARI, N. Y PALMER, M. (1994). *Current issues in computational linguistics: In honour of Don Walker*. Ed. Springer Science & Business Media.

Anexo 1. Cuantificadores

ADJETIVOS CUANTIFICADORES		MASCULINO	FEMENINO
TOTALES	**SINGULAR**	cada, todo	cada, toda
	PLURAL	ambos, sendos, todos	ambas, sendas, todas
PARCIALES	**SINGULAR**	algotro, algún, alguno, bastante, cualquier, cualquiera, cuánto, cuanto, demasiado, más, menos, mucho, poco, tanto	algotra, alguna, bastante, cualquier, cualquiera, cuánta, cuanta, demasiada, más, menos, mucha, poca, tanta
	PLURAL	algotros, algunos, bastantes, cuántos, cuantos, demasiados, más, menos, muchos, pocos, tantos, varios	algotras, algunas, bastantes, cuántas, cuantas, demasiadas, más, menos, muchas, pocas, tantas, varias
NULOS	**SINGULAR**	ningún, ninguno	ninguna
	PLURAL	ningunos	ningunas

PRONOMBRES CUANTIFICADORES		MASCULINO	FEMENINO	NEUTRO
TOTALES	**SINGULAR**	todo	toda	todo
	PLURAL	ambos, todas	ambas, todas	
PARCIALES	**SINGULAR**	algotro, alguien, alguno, bastante, cualquiera, cuánto, cuanto, demasiado, más, menos, mucho, poco, quienquiera, tanto	algotra, alguien, alguna, bastante, cualquiera, cuánta, cuanta, demasiada, más, menos, mucha, poca, quienquiera, tanta	algo, cuanto, mucho, poco, tanto bastante, demasiado, más, menos
	PLURAL	algotros, algunos, bastantes, cualesquiera, cuántos, cuantos, demasiados, más, menos, muchos, pocos, quienesquiera, tantos, varios	algotras, algunas, bastantes, cualesquiera, cuántas, cuantas, demasiadas, más, menos, muchas, pocas, quienesquiera, tantas, varias	
NULOS	**SINGULAR**	ninguno, nadie	ninguna, nadie	nada
	PLURAL	ningunos	ningunas	

ADVERBIOS	**TOTALES**	todo
	PARCIALES	algo, bastante, cuan, cuánto, cuanto, demasiado, más, menos, mucho, muy, poco, tan, tanto
	NULOS	nada

Anexo 2. Numerales

CARDINALES	valor cardinal cuantificador	ADJ	PN	N	
	valor ordinal identificador	ADJ	PN	N	
ORDINALES	valor ordinal identificador	ADJ	PN	N	ADV
FRACCIONARIOS	valor partitivo cuantificador	ADJ	PN	N	ADV
MULTIPLICATIVOS	valor multiplicativo cuantificador	ADJ		N	

ADJETIVOS CARDINALES		**MASCULINO**	**FEMENINO**
VALOR CARDINAL DISTRIBUCIÓN PRENOMINAL	**SINGULAR**	un	una
		cero	cero
	PLURAL	cero	cero
		dos…	dos…
		veinte	veinte
		veintiún	veintiuna
		veintidós…	veintidós…
		treintaiún…	treintaiuna…
		cuarenta…	cuarenta…
		noventainueve	noventainueve
		cien	cien
		doscientos…	doscientas…
		mil	mil
VALOR ORDINAL DISTRIBUCIÓN POSNOMINAL	**SINGULAR**	uno	una
	PLURAL	dos	dos
		veintiuno	veintiuna
		veintidós…	veintidós…
		treintaiuno…	treintaiuna…
		cuarenta…	cuarenta…
		noventainueve	noventainueve
		cien…	cien…
		doscientos…	doscientas…
		mil	mil

SUSTANTIVOS CARDINALES LAS CIFRAS	SINGULAR	PLURAL
MASCULINO	cero	ceros
	uno	unos
	dos…	doses…
	diez…	dieces…
	veintiuno	veintiunos
	treintaiuno...	treintaiunos…
	noventainueve	noventainueves
	cien	cienes
	ciento	cientos
	doscientos…	doscientos…
	mil, millar	miles, millares
	millón	millones
	millardo	millardos
	billón	billones
	trillón	trillones
	cuatrillón	cuatrillones
	quintillón	quintillones
	sextillón	sextillones
	septillón	septillones
	octillón	octillones
	nonillón	nonillones
	decillón	decillones
	undecillón	undecillones
	duodecillón	duodecillones
	tredecillón	tredecillones
	cuatordecillón	cuatordecillones
	quindecillón	quindecillones
	sexdecillón	sexdecillones
	septendecillón	septendecillones
	octodecillón	octodecillones
	novendecillón	novendecillones
	vigintillón	vigintillones

SUSTANTIVOS CARDINALES LAS HORAS	SINGULAR	PLURAL
FEMENINO	una	
		cero
		dos
		tres
		cuatro…

PRONOMBRES CARDINALES	MASCULINO	FEMENINO
SINGULAR	cero	cero
	uno	una
PLURAL	cero	cero
	dos	dos
	veintiuno…	veintiuna…
	noventainueve	noventainueve
	cien	cien
	doscientos…	doscientas…
	mil	mil

ADJETIVOS ORDINALES	MASCULINO	FEMENINO
SINGULAR	primer, primero	primera
	segundo	segunda
	tercer, tercero	tercera
	cuarto	cuarta
	quinto	quinta
	sexto	sexta
	séptimo, sétimo	séptima, sétima
	octavo, ochavo	octava, ochava
	noveno, nono	novena, nona
	décimo	décima
	decimoprimer, decimoprimero, undécimo	decimoprimera, undécima
	duodécimo, decimosegundo	duodécima, decimosegunda
	tredécimo, decimotercio decimotercer, decimotercero…	tredécimo, decimotercio, decimotercera…
	vicésimo, veintésimo, vigésimo, vigesimoprimer, vigesimoprimero…	vicésima, veintésima, vigésima, vigesimoprimera…

	tricésimo,trigésimo, trecésimo…	tricésima, trigésima, trecésima…
	nonagesimonono	nonagesimonona
	centeno	centena
	centésimo	centésima
	ducentésimo…	ducentésima…
	milésimo	milésima
	dosmilésimo…	dosmilésima…
	cienmilésimo	cienmilésima
	quinientosmilésimo	quinientosmilésima
	millonésimo	millonésima
	diezmillonésimo	diezmillonésima
	cienmillonésimo	cienmillonésima
	milmillonésimo	milmillonésima
	diezmilmillonésimo	diezmilmillonésima
	cienmilmillonésimo	cienmilmillonésima
	billonésimo	billonésima
	trillonésimo…	trillonésima…
PLURAL	primeros	primeras
	segundos	segundas
	terceros	terceras
	cuartos	cuartas
	quintos	quintos
	sextos	sextos
	séptimos, sétimos	séptimas, sétimas
	octavos, ochavos	octavas, ochavas
	novenos, nonos	novenas, nonas
	décimos	décimas
	decimoprimeros, undécimos	decimoprimeras, undécimas
	duodécimos, decimosegundos	duodécimas, decimosegundas
	tredécimos, decimotercios, decimoterceros…	tredécimas, decimotercias, decimoterceras…
	vicésimos, veintésimos, vigésimos, vigesimoprimeros…	vicésimas, veintésimas, vigésimas, vigesimoprimeras…
	tricésimos, trigésimos, trecésimos…	tricésimas, trigésimas, trecésimas…
	nonagesimononos	nonagesimononas
	centenos	centenas
	centésimos	centésimas
	ducentésimos…	ducentésimas…
	milésimos	milésimas

	dosmilésimos…	dosmilésimas…
	cienmilésimos	cienmilésimas
	quinientosmilésimos	quinientosmilésimas
	millonésimos	millonésimas
	diezmillonésimos	diezmillonésimas
	cienmillonésimos	cienmillonésimas
	milmillonésimos	milmillonésimas
	diezmilmillonésimos	diezmilmillonésimas
	cienmilmillonésimos	cienmilmillonésimas
	billonésimos	billonésimas
	trillonésimos…	trillonésimas…

PRONOMBRES ORDINALES	**MASCULINO**	**FEMENINO**
SINGULAR	primero	Primera
	segundo	Segunda
	tercero	Tercera
	Cuarto	Cuarta
	Quinto	Quinta
	Sexto	Sexta
	séptimo, sétimo	séptima, sétima
	octavo, ochavo	octava, ochava
	noveno, nono	novena, nona
	Décimo	Décima
	decimoprimero, undécimo	decimoprimero, undécima
	duodécimo, decimosegundo	duodécima, decimosegunda
	tredécimo, decimotercio decimotercero…	tredécimo, decimotercio, decimotercera,
	vicésimo, veintésimo, vigésimo, vigesimoprimero…	vicésima, veintésima, vigésima, vigesimoprimera…
	tricésimo, trigésimo, trecésimo…	tricésima, trigésima, trecésima…
	nonagesimonono	Nonagesimonona
	centeno	Centena
	centésimo	Centésima
	ducentésimo…	ducentésima…
	milésimo	Milésima
	dosmilésimo…	dosmilésima…
	cienmilésimo	Cienmilésima

	quinientosmilésimo	Quinientosmilésima
	millonésimo	Millonésima
	diezmillonésimo	Diezmillonésima
	cienmillonésimo	Cienmillonésima
	milmillonésimo	Milmillonésima
	diezmilmillonésimo	Diezmilmillonésima
	cienmilmillonésimo	Cienmilmillonésima
	billonésimo	Billonésima
	trillonésimo…	trillonésima…
PLURAL	primeros	Primeras
	segundos	Segundas
	terceros	Terceras
	cuartos	Cuartas
	quintos	Quintos
	sextos	Sextos
	séptimos, sétimos	séptimas, sétimas
	octavos, ochavos	octavas, ochavas
	novenos, nonos	novenas, nonas
	décimos	Décimas
	decimoprimeros, undécimos	decimoprimeras, undécimas
	duodécimos, decimosegundos	duodécimas, decimosegundas
	tredécimos, decimotercios, decimoterceros…	tredécimas, decimotercias, decimoterceras…
	vicésimos, veintésimos, vigésimos, vigesimoprimeros…	vicésimas, veintésimas, vigésimas, vigesimoprimeras…
	tricésimos, trigésimos, trecésimos…	tricésimas, trigésimas, trecésimas…
	nonagesimononos	Nonagesimononas
	centenos	Centenas
	centésimos	Centésimas
	ducentésimos…	ducentésimas…
	milésimos	Milésimas
	dosmilésimos…	dosmilésimas…
	cienmilésimos	Cienmilésimas
	quinientosmilésimos	Quinientosmilésimas
	millonésimos	millonésimas
	diezmillonésimos	diezmillonésimas
	cienmillonésimos	cienmillonésimas
	milmillonésimos	milmillonésimas
	diezmilmillonésimos	diezmilmillonésimas

	cienmilmillonésimos	cienmilmillonésimas
	billonésimos	billonésimas
	trillonésimos…	trillonésimas…

SUSTANTIVOS ORDINALES	MASCULINO	FEMENINO
SINGULAR	primero	primera
	segundo	segunda
	tercero	tercera
	cuarto	cuarta
	quinto	quinta
	sexto	sexta
	séptimo, sétimo	séptima, sétima
	octavo, ochavo	octava, ochava
	noveno, nono	novena, nona
	décimo…	décima…
PLURAL	primeros	primeras
	segundos	segundas
	terceros	terceras
	cuartos	cuartas
	quintos	quintas
	sextos	sextas
	séptimos, sétimos	séptimas, sétimas
	octavos, ochavos	octavas, ochavas
	novenos, nonos	novenas, nonas
	décimos…	décimas…

ADJETIVOS FRACCIONARIOS	MASCULINO	FEMENINO
SINGULAR	medio	media
		tercera
		cuarta
		quinta…
		décima
		undécima, onceava, onzava
		duodécima, doceava, dozava…
		vigésima, vicésima, veintava, veinteava, veintésima…
		trecésima, treintava, tricésima, trigésima
		treintaiunava…
		centava, centésima, céntima…

		milésima…
		millonésima…
PLURAL	medios	Medias
		terceras
		cuartas
		quintas…
		décimas
		undécimas, onceavas, onzavas
		duodécimas, doceavas, dozavas…
		vigésimas, vicésimas, veintavas, veintésimas…
		trecésimas, treintavas, tricésimas, trigésimas
		treintaiunavas…

PRONOMBRES FRACCIONARIOS	**MASCULINO**	**FEMENINO**
SINGULAR	medio	media
		tercera
		cuarta
		quinta…
		décima
		undécima, onceava, onzava
		duodécima, doceava, dozava…
		vigésima, vicésima, veintava, veinteava, veintésima…
		trecésima, treintava, tricésima, trigésima
		treintaiunava…
		centava, centésima, céntima…
		milésima…
		millonésima…
PLURAL	medios	medias
		terceras
		cuartas
		quintas…
		décimas
		undécimas, onceavas, onzavas
		duodécimas, doceavas, dozavas…
		vigésimas, vicésimas, veintavas, veintésimas…
		trecésimas, treintavas, tricésimas, trigésimas
		treintaiunavas…

ADVERBIOS ORDINALES	primero, segundo, tercero, cuarto…

SUSTANTIVOS FRACCIONARIOS	MASCULINO	FEMENINO
SINGULAR	medio	media, mitad
	tercio	
	cuarto	
	quinto	
	sexto	
	séptimo	
	octavo	
	noveno	
	décimo	
	undécimo, onceavo	
	duodécimo, doceavo	
	vigésimo, veintavo, veinteavo	
	trigésimo, treintavo	
	centésimo	centésima
	milésimo	milésima
	millonésimo	millonésima
PLURAL	medios	medias, mitades
	tercios	
	cuartos	
	quintos	
	sextos	
	séptimos	
	octavos	
	novenos	
	décimos	
	undécimos, onceavos	
	duodécimos, doceavos...	
	vigésimos, veintavos, veinteavos…	
	trigésimos, treintavos…	
	centésimos	centésimas
	milésimos	milésimas
	millonésimos	millonésimas

ADJETIVOS MULTIPLICATVOS	MASCULINO	FEMENINO
SINGULAR	doble, dúplice	doble, dúplice
	triple, tríplice, triplo	triple, tríplice, tripla
	cuádruple cuádruplo	cuádruple, cuádrupla
	quíntuple, quíntuplo	quíntuple, quíntupla
	séxtuple, séxtuplo	séptuple, séxtupla
	séptuple, séptuplo	séptuple, séptupla
	óctuple, óctuplo	óctuple, óctupla
	nónuplo	nónupla
	céntuplo	céntupla
PLURAL	dobles	doble
	triples, triplos	triples, triplas
	cuádruples cuádruplos	cuádruples, cuádruplas
	quíntuples, quíntuplos	quíntuples, quíntuplas
	séxtuples, séxtuplos	séptuples, séxtuplas
	séptuples, séptuplos	séptuples, séptuplas
	óctuples, óctuplos	óctuples, óctuplas
	nónuplos	nónuplas
	céntuplos	céntuplas

SUSTANTIVOS MULTIPLICATIVOS	MASCULINO
SINGULAR	doble, dúplice
	triple, tríplice, triplo
	cuádruple cuádruplo
	quíntuple, quíntuplo
	séxtuple, séxtuplo
	séptuple, séptuplo
	óctuple, óctuplo
	nónuplo
	céntuplo
PLURAL	dobles
	triples, triplos
	cuádruples cuádruplos
	quíntuples, quíntuplos
	séxtuples, séxtuplos
	séptuples, séptuplos
	óctuples, óctuplos
	nónuplos
	céntuplos

Anexo 3. Indefinidos

		MASCULINO	FEMENINO
ADJETIVOS	SINGULAR	cierto, demás, otro	cierta, demás, otra,
	PLURAL	ciertos, demás otros	ciertas, demás, otras
PRONOMBRES	SINGULAR	otro, uno	otra, una
	PLURAL	otros, unos	otra, una
ARTÍCULOS	SINGULAR	un	una
	PLURAL	unos	unas

Anexo 4. Posesivos

ADJETIVOS POSESIVOS		MASCULINO	FEMENINO
		PRIMERA PERSONA	
UN POSEEDOR	UN OBJETO	mi, mío	mi, mía
	VARIOS OBJETOS	mis, míos	mis, mías
VARIOS POSEEDORES	UN OBJETO	nuestro	nuestra
	VARIOS OBJETOS	nuestros	nuestras
		SEGUNDA PERSONA	
UN POSEEDOR	UN OBJETO	tu, tuyo, su, suyo	tu, tuya, su, suya
	VARIOS OBJETOS	tus, tuyos, su, suyos	tus, tuyas, sus, suyas
VARIOS POSEEDORES	UN OBJETO	vuestro	vuestra
	VARIOS OBJETOS	vuestros	vuestras
		TERCERA PERSONA	
UN POSEEDOR	UN OBJETO	su, suyo	su, suya
	VARIOS OBJETOS	sus, suyos	sus, suyas
VARIOS POSEEDORES	UN OBJETO	su	su
	VARIOS OBJETOS	sus	sus

PRONOMBRES POSESIVOS		MASCULINO	FEMENINO
		PRIMERA PERSONA	
UN POSEEDOR	UN OBJETO	mío	mía
	VARIOS OBJETOS	míos	mías
VARIOS POSEEDORES	UN OBJETO	nuestro	nuestra
	VARIOS OBJETOS	nuestros	nuestras
		SEGUNDA PERSONA	

UN POSEEDOR	**UN OBJETO**	tuyo, suyo	tuya, suya
	VARIOS OBJETOS	tuyos, suyos	tuyas, suyas
VARIOS POSEEDORES	**UN OBJETO**	vuestro	vuestra
	VARIOS OBJETOS	vuestros	vuestras
		TERCERA PERSONA	
UN POSEEDOR	**UN OBJETO**	suyo	suya
	VARIOS OBJETOS	suyos	suyas
VARIOS POSEEDORES	**UN OBJETO**	su	su
	VARIOS OBJETOS	sus	sus

Anexo 5. Relativos

ADJETIVOS INTERROGATIVOS	MASCULINO	FEMENINO
SINGULAR	cuyo, cuanto	cuya, cuanta
PLURAL	cuyos, cuantos	cuyas, cuantas

PRONOMBRES INTERROGATIVOS	MASCULINO	FEMENINO	NEUTRO
SINGULAR	que, quien, cuanto, cual	que, quien, cuanta, cual	que cuanto
PLURAL	que, quienes, cuantos, cuales	que, quienes, cuantas, cuales	

ADVERBIOS INTERROGATIVOS	cuando, como, donde, adonde, cuanto

Anexo 6. Interrogativos y exclamativos

ADJETIVOS INTERROGATIVOS	MASCULINO	FEMENINO
SINGULAR	qué, cuál, cuánto	qué, cuál, cuánta
PLURAL	qué, cuáles, cuántos	qué, cuáles, cuántas

PRONOMBRES INTERROGATIVOS	MASCULINO	FEMENINO	NEUTRO
SINGULAR	cuál, cuánto, qué, quién	cuál, cuánta, qué, quién	qué cuánto
PLURAL	cuáles, cuántos, qué, quiénes	cuáles, cuántas, qué, quiénes	

ADVERBIOS INTERROGATIVOS	cuándo, cómo, dónde, cuánto

Anexo 7. Demostrativos

ADJETIVOS DEMOSTRATIVOS	MASCULINO	FEMENINO
SINGULAR	este ese aquel tal semejante	esta esa aquella tal semejante
PLURAL	estos esos aquellos tales semejantes	estas esas aquellas tales semejantes

PRONOMBRES DEMOSTRATIVOS	MASCULINO	FEMENINO	NEUTRO
SINGULAR	este ese aquel tal	esta esa aquella tal	esto eso aquello
PLURAL	estos esos aquellos tales	estas esas aquellas tales	

ADVERBIOS DEMOSTRATIVOS	aquí, ahí, allí, acá, allá, ahora, hoy, así, entonces

Anexo 8. Repertorio de listas

1. Sustantivos femeninos que empiezan con á-, *há-*
2. Verbos
3. Siglas
4. Abreviaturas
5. Extranjerismos crudos
6. Infinitivos
7. Adjetivos de grado extremo
8. Monosílabos que se escriben sin tilde
9. Adverbios de grado
10. *Pluralia tantum* en singular
11. *Singularia tantum* en plural
12. Locuciones latinas y latinismos
13. Sustantivos colectivos
14. Palabras acabadas en *-í*
15. Palabras acabadas en *-ú*
16. Gerundios
17. Participios
18. Imperativos
19. Verbos pronominales
20. Conectores adverbiales
21. Palabras que terminan en *-y* formando diptongo o triptongo
22. Extranjerismos que terminan con la secuencia «consonante + y»
23. Palabras sin flexión para plural
24. Nombres compuestos por «vb+N» incontables siempre en singular.
25. Nombres compuestos por «vb+N» contables siempre en plural.
26. Numerales
27. Cardinales
28. Ordinales
29. Fraccionarios
30. Multiplicativos
31. Segunda persona del singular del pretérito indefinido de indicativo de todos los verbos
32. Palabras que empiezan por *espl-*
33. Sustantivos que se refieren a partes del cuerpo
34. Sustantivos abstractos
35. 2ª persona del singular del pretérito indefinido de indicativo de todos los verbos+*s.* Formas erróneas.
36. Expresiones temporales con valor de pasado
37. Sustantivos terminados en «vocal+*y*»

Anexo 9. Abreviaturas

- det
- art
- cuantif
- num; número/ letra
- sg
- pl
- masc
- fem
- superl
- N
- propio
- adj
- adv
- prep
- nx
- conj
- coord: coordinada/coordinante
- subord: subordinada/subordinante
- pn
- pn suj
- pn CD
- pn CI
- pn reflex (compuesto de pn de CD y pn de CI)
- vb
- inf
- participio
- ger
- ind
- subj
- imperat
- pres
- pret
- indef
- perf
- imperf
- simp
- comp
- plusc
- fut
- cond

Anexo 10. Incidencias

INCIDENCIAS	
ID	**Incidencia**
0	ERRORES ofrece corrección
1	Metátesis: (l) *cluquillas*.
2	Epéntesis.
3	Ortografía: error formal por asimilación con otras pautas de composición (*convezca*).
4	Ortografía: uso de la *ex* y *es-*
5	Tipografía: segmentación de las palabras
6	Ortografía: uso de la hache intercalada.
7	Ortografía: tilde en palabras llanas que acaban en doble consonante aunque la consonante final sea *-s* o *-n*.
8	Ortografía: los monosílabos no llevan tilde.
9	Ortografía: tilde diacrítica.
10	Tipografía: Mayúsculas: se escribe con minúscula los nombres comunes que designan días de la semana, meses y estaciones del año.
12	Morfología prefijación: (*EX-*) derivadas con el prefijo "*EX*"
13	Palabra incompleta
14	Tipografía: Mayúsculas: acentuación de las vocales mayúsculas.
15	Tipografía: Puntuación (,) : detrás de "*pero*" no se escribe coma.
16	Tipografía: Puntuación (.) : la oración entre paréntesis no se cierra con punto.
17	Extranjerismos: *k>qu e,i/c a, o, u*.
18	Tipografía: Puntuación (,) : la coma no se puede poner delante de los tres puntos suspensivos.
19	Tipografía: Puntuación (…) : no se puede combinar el etcétera y los tres puntos suspensivos.
20	Tipografía: Numerales: no se puede separar con un punto los millares por confusión con los decimales.
21	Tipografía: Numerales: no se puede separar un número decimal mediante un apóstrofo.
22	Tipografía: Horas: no se puede escribir símbolo *h*, "*horas*" pegada al último dígito
23	Tipografía: Abreviaturas: no se puede suprimir el punto de las abreviaturas porque es lo que señala que una palabra está cortada.
24	Tipografía: Abreviaturas: tras el punto de las abreviaturas no se podrá escribir un punto final.

INCIDENCIAS	
25	Tipografía: Abreviaturas: si la palabra completa lleva tilde y en la abreviatura incluye esa vocal sobre la que cae el acento, debe mantenerse la tilde. No se puede romper el dígrafo compuesto por las dos erres.
26	Tipografía: Abreviaturas: la abreviatura de siglo es " *s.* ", en minúscula.
27	Tipografía: Mayúsculas: artículo que forma parte del topónimo va en mayúscula.
28	Tipografía: Mayúsculas: la "*U*" mayúscula se escribir con diéresis(¨) cuando lo exija su pronunciación.
29	Tipografía: Abreviaturas: cuando la abreviatura tenga letra voladita si habrá que escribir punto antes de la letra voladita y puede llevar punto final.
30	Tipografía: Puntuación (,) : las construcciones comparativas no se separan por comas.
31	Tipografía: Numerales: el símbolo " % " va acompañado de la cifra escrita en números
32	*Ahí-hay*: vacilación entre formas; *hay/ay por ahí*
33	*Ahí-hay*: vacilación entre formas; *ahí/ay por hay*
34	Error de construcción en una expresión o forma: *antes de mí*
36	Ortografía: palabra sin tilde
37	Concordancia suj-vb: casos especiales; pers oraciones escindidas…..
38	Ortografía: expresiones con "*hecho*"- "*echo*"
39	Ortografía: relativos inespecíficos (*adondequiera que*)
40	Ortografía: *conque / con que*
41	Morfología: construcción de plurales
42	Ortografía: sustantivación con el artículo
43	Ortografía: tilde diacrítica *CUAL(ES)*
44	Ortografía: tilde diacrítica *QUIEN(ES)*
45	Ortografía: tilde diacrítica *CUANDO*
46	Tipografía: Puntuación (:) : dos puntos y las preposiciones
47	Ortografía: tilde diacrítica *CUANTO/A,CUANTOS/AS*
48	Ortografía: tilde diacrítica *QUE*.
49	Ortografía: tilde diacrítica *COMO*.
50	Léxico: neologismo innecesario no recogido en DRAE. Se ofrece una alternativa admitida por la norma.
51	Morfología verbal: error en construcción de una forma vbl o sus pn.
56	Semántica: confusión en sentido o régimen de una palabra (dimitido-dimisionario)
58	Concordancia: persona
68	Homófonos: Deshecho: *deshecho*
501	Semántica: restricciones semánticas (*rechace*)

INCIDENCIAS	
502	Morfología: no cumple con las reglas de composición morfológica; Vb+N
503	Doble opción: una es preferida por la norma culta y registro escrito (*más nunca*), *mayoría de* + art
504	Concordancia: género (*la detective privada*)
505	Concordancia: género, excepción (*la ama de llaves, mucho agua*)
506	Concordancia: número
507	Concordancia gramatical: categorías: *mejores preparados*
508	Expresiones latinas: ortografía y expresión correcta
509	Error de régimen preposicional: este verbo no rige preposición
510	Error de régimen preposicional: este verbo rige una preposición diferente
511	Error de régimen preposicional: este término o expresión rige una preposición diferente
512	Error por redundancia; *muy abominable*, *subir para arriba*
513	Semántica: error en una expresión por confusión/interferencia con otro término, (*cuanto más, cuando más*...) o inclusión de prep (*arriba a abajo*)...
514	Sintaxis de "*antes*" y "*después*"
516	Léxico: palabra o expresión coloquial, no admitida en lengua escrita ni formal.
518	Léxico-semántico: forma con diferente significado según su grafía (compacta/segmentado) (*demás-de más*)
520	Ortografía: error en homófonos (*iba-iva*)
521	*asímismo*: Ortografía: error
522	Ortografía: error entre casi homófonos; *arroyo-arrollo*,
523	Casi homófonos: *haya, halla*
530	Error de número: la palabra es siempre pl (incluye adv invariables)
531	Error de número: la palabra es siempre sg (incluye adv invariables)
532	Gramática verbal: uso del gerundio
533	Error de género: la palabra es fem.
540	Léxico-semántico: expresión o palabra en desuso.
550	Dequeísmo: dequeísmo en régimen verbal general
551	Dequeísmo: dequeísmo en locución o expresión
552	Queísmo: supresión indebida de "*de*" en rección verbal general
553	Queísmo: supresión indebida de "*de*" en expresión o locución
554	Queísmo; supresión de otra prep diferente de "*de*"
555	Ortografía: errores frecuentes, falta competencia
560	Haber y hay: uso indebido del terciopersonal
570	Semántica: incoherencia; *bajo la base de, destornillarse de risa*
571	Léxico: adaptación de un extranjerismo

INCIDENCIAS	
572	Léxico: extranjerismo. Se prefiere la versión española
588	Concordancia gramatical: suj-vb
700	*SI NO, SINO*
701	*sino/si no:* vacilación
800	*POR QUÉ, PORQUE, PORQUÉ*... (OLE 558)
801	*Por qué* y sus variantes. Anteposición indebida del artículo
802	*por qué- porque*: Vacilación de formas
815	Gramática verbal: vacilación entre tiempos o modos del verbo
816	Gramática verbal: impersonales
990	Morfología: plural de latinismos y helenismos
1000	ERRORES GENERALES- ofrece corrección
1001	Tipografía: Mayúsculas: tras el signo de cierre de exclamación se escribe con mayúscula.
1002	Tipografía: Mayúsculas: tras el signo de cierre de interrogación se escribe con mayúscula.
1004	Tipografía: Puntuación (;) : tras el punto y coma no se escribe mayúscula, salvo si es nombre propio.
1005	Ortografía: usos de las conjunciones "o" e "y"
1007	Tipografía: Abreviaturas: las abreviaturas no se pueden cortar cuando no caben enteras en un renglón
1008	Tipografía: Numerales: las expresiones numéricas en cifras no se pueden cortar cuando no caben enteras en un renglón.
1009	Ortografía: palabras que están separadas con barra no se pueden cortar cuando no caben enteras en un renglón.
1010	Tipografía: Puntuación (,) : los adverbios conectores en medio de la frase van entre comas.
1011	Tipografía: Puntuación (,) : conectores al principio y al final de la frase se deben aislar de esta mediante comas
1012	Estilo: Numerales: la locución " por ciento" se utiliza cuando hablamos de porcentajes donde la cifra está escrita con letras.
1013	Morfología prefijación: (*EX*-) derivadas con el prefijo "*EX*"
1014	Concordancia: N femeninos que empiezan por a/ha tónica.
1015	Distinción de género: Ciudadanos y ciudadanas- ciudadanos
1018	Morfología verbal: epéntesis en 2º pers sg de pretérito perfecto de indicativo.
1019	Tipografía: Extranjerismos: las palabras extranjeras no pueden cortarse al final de renglón
1020	Tipografía: Puntuación (¿¡!?) : Orden. Se puede combinar los signos de exclamación e interrogación siempre y cuando se respete el mismo orden para la apertura y el cierre.

INCIDENCIAS	
1021	Tipografía: Puntuación (…) : tras los puntos suspensivos se puede escribir cualquier signo de puntuación, excepto el punto.
1022	Ortografía i/y: extranjerismos acabados en “consonante + -y “al final palabra se adaptan al español cambiando la “–y” por "i"
1023	Ortografía i/y: fonema “–y” átono en posición final precedido de una o dos vocales con las que forma un diptongo o triptongo.
1024	Morfología prefijación: palabras prefijadas se escriben como una sola unidad léxica
1025	Morfología prefijación: siglas, nombres propios univerbales y números.
1026	Ortografía -y: palabras llanas que forman un diptongo acabadas en -y llevan tilde.
1050	Tipografía: Siglas: se escriben sin punto final
1051	Siglas: se escriben sin tilde
1052	Siglas: se escriben sin plural
1053	Tipografía: Siglas: no se pueden cortar cuando no caben enteras en un renglón
1054	Tipografía: Siglas: se escriben con mayúsculas
1060	Morfología: palabra sin flexión de pl. (test)
1100	PROBLEMAS CON A, AH, HA, E, EH, HE
1101	Ortografía: "ha" es la tercera persona del singular del pret. indicativo del verbo haber y su uso actual es de verbo auxiliar en formas compuestas, va seguido siempre de un participio.
1102	Ortografía:"ha" en la perífrasis verbal: "ha" + de + verbo en infinitivo.
1103	Ortografía: "he" es la primera persona del singular del pret. indicativo del verbo haber y su uso actual es de verbo auxiliar en formas compuestas, va seguido siempre de un participio.
1107	Concordancia: persona (ustedes os calláis)
1501	Morfología verbal: epéntesis en construcción del imperativo
1502	Morfología verbal: error en la construcción de formas enclíticas
1660	Tipografía: extranjerismos y locuciones latinas
2000	RECOMENDACIÓN-ofrece alternativa, no automática
2001	Doble opción: concordancia
2003	Tipografía: Grafía: una opción es más recomendada que la otra
2004	Léxico-Semántico: forma con mismo significado y diferente grafía (compacta/segmentada). Ofrecemos la más recomendada
2005	Tipografía: Numerales: la escritura de las fracciones en cifras así como de los números fuera de contextos matemáticos o técnicos deben escribirse con letras.
2006	Tipografía: Numerales: no se puede escribir las dos últimas cifras del año con apóstrofe
2007	Tipografía: Horas: escribir las horas con letras señalando la franja horaria

INCIDENCIAS	
2008	Estilo: Numerales: Preferencia de multiplicativos acabados en -e
2010	Semántica: palabra o expresión desaconsejada por motivos semánticos (en cuyo caso)
2011	Ortografía: doble opción en ortografía, una recomendada por la RAE.
2501	Doble opción según variedad geográfica España-América
2502	Léxico: restos de antiguos vocablos (desque), vocablos en desuso. Se ofrece una alternativa admitida por la norma.
2503	Léxico: vulgarismo
2504	Semántica: incoherencia (antiedad, embarazado). No está en el DRAE
2505	Doble opción por causa de género: una de ellas más recomendable (polígloto/a, dote, mimbre).
2507	Léxico: término no recomendado en lengua escrita (qué tonto que es)
2508	Expresión desaconsejada por la norma culta (a día de hoy),
2509	Semántica: restricciones semánticas (inicializar)
2510	Expresión o palabra desaconsejada en lenguaje formal. Ofrece otra opción.
2511	Léxico: expresión derivada de lengua extranjera. Se prefiere la versión española
3000	RECOMENDACIÓN GENERALIZADA- ofrece corrección, no automática
3001	Morfología prefijación: casos de prefijación donde es frecuente la coincidencia de vocales iguales y esto produce formas con vocal doble o reducida.
3002	Extranjerismos: k >qu e,i / c a, o, u.
3210	Doble opción: (una de ellas recomendada por la norma). Concordancia
3511	Léxico: expresión derivada de lengua extranjera en términos estructurales. Se prefiere la versión española
4000	AVISOS LING- no ofrece corrección automática
4001	Acentuación: expresión contiene el det.posesivo "tu" y debe aparecer seguido de un nombre.
4002	Ortografía: expresión que contiene un monosílabo con tilde diacrítica para diferenciarlo de otra forma idéntica pero con diferente función.
4003	Léxico: la palabra no existe; se ofrece la alternativa más viable.
4004	Ortográfico: Tilde diacrítica
4005	Error de número: la palabra es siempre pl (incluye adv invariables)
4006	Morfología: posible error en flexión (as-aes)
4007	Tipografía: Puntuación (,) : "pero" va entre comas cuando aparece un vocativo o una interjección: ,pero,
4009	Tipografía: Puntuación (,) : cuando los adverbios conectores se sitúan en medio de la frase es optativo ponerlos entre comas si antes tienen un nexo subordinante o una conjunción coordinante.
4010	Léxico-Semántico: precisión de algunos dobletes o términos que parecen sinónimos: islamista/islámico

INCIDENCIAS	
4011	Tipografía: Horas: si se trata de horas en punto podemos prescindir de los últimos dígitos, pero el símbolo h sería obligatoria en este caso
4012	Léxico-Semántico: impropiedades léxicas
4013	Léxico-Semántico: forma con diferente significado según su grafía (compacta/segmentada)
4014	Léxico-Semántico: forma con mismo significado y diferente grafía (compacta/segmentada)
4015	Porque: usos de la forma
4019	Ahí, ay, hay: usos de las formas
4020	Léxico: diferencia entre homófonas
4021	Léxico: diferencia entre casi homófonas; esotérico-exotérico
4022	Ortografía: TILDE EN LOS PRONOMBRES INTERROGATIVOS Y EXCLAMATIVOS.
4023	Tipografía: Abreviaturas: la única abreviatura que no se escribe con punto y se enmarca con paréntesis doble es alias (a) pero se prefiere la palabra completa
4024	Concordancia: posible error
4025	Ortografía: hecho/echo
4030	Error gramatical: construcción no admitida
4041	Posible error de concordancia: mejores preparados, media mareada,
4500	Semántica: restricciones semánticas (barajar, mayoría + N colectivo o pl)
4502	Gramática verbal: régimen preposicional
4503	Semántica: error en una expresión por confusión/interferencia con otro término, (cuanto más, cuando más...)
4504	Morfología verbal: vacilación entre tiempos o modos verbales; venir-venid
4505	Ortografía: con que, con que
4506	Ortografía: con (artículo) que
4507	Ortografía: con (artículo) cual, el "que" se puede sustituir por otro relativo equivalente "cual".
4508	Expresión o palabra desaconsejada en lenguaje formal
4509	Cacofonía: necesita revisión
4511	Léxico: posible error en la palabra
4520	SINO-SI NO
4521	sino: Vacilación, en ausencia de comas, de la conj adversativa, "No sé si no vendrá- No hago sino quererte"
4530	ASÍ MISMO, ASIMISMO
4531	Vacilación entre formas
4540	Latinismos: usos
4600	Error entre casi homófonos; haya, halla, aya

INCIDENCIAS	
5000	AVISO GENERALIZADO- no ofrece corrección automática
5001	Morfología verbal: posible vacilación entre formas o construcciones verbales
5501	Semántica: precisión en palabra o expresión.
5505	Gramática verbal: vacilación entre tiempos o modos del verbo
5510	Gramática verbal: uso del infinitivo

Anexo 11. Fenómenos asociados

FENÓMENOS
Concordancia género: excepciones.
Concordancia gramatical: categorías
Concordancia gramatical: suj-vb.
Concordancia: género.
Concordancia: número.
Concordancia: persona
Construcción de una expresión.
Construcción de una palabra.
Doble opción: una es preferible en la lengua actual y recomendada por la norma culta.
Estilo: Redundancia.
Expresión o forma errónea solo parcialmente. En algunas zonas o contextos está aceptada.
Expresiones latinas.
Extranjerismos: palabra o expresión
Género.
Gramática verbal: general.
Gramática verbal: impersonalidad.
Gramática verbal: usos del gerundio.
Gramática verbal: usos del infinitivo.
Gramática: Dequeísmo.
Gramática: Queísmo.
Gramática: Régimen preposicional.
Léxico: arcaísmo.
Léxico: neologismo.
Léxico: palabra lexicalizada como nombre común.
Léxico: vulgarismo.
Léxico-Estilo: coloquialismo o lengua oral
Léxico-Semántico: impropiedades léxicas.
Léxico-Semántico: palabras que se pueden escribir juntas o separadas pero cambian su significado.
Léxico-Semántico: palabras que se pueden escribir juntas o separadas sin cambio de significado.

FENÓMENOS
Léxico-Semántico: término o expresión en desuso o desaconsejado por la norma en el registro escrito.
Morfología verbal.
Morfología: construcción de plurales de latinismos
Morfología: construcción de plurales.
Morfología: general.
Morfología: prefijación.
Morfología: reglas de composición de N.
Morfología: sufijación.
Numerales.
Número.
Ortografía: Acentuación.
Ortografía: Acentuación: monosílabos.
Ortografía: b/v.
Ortografía: general.
Ortografía: palabras que admiten dos formas ortográficamente distintas.
Ortografía: preposiciones.
Ortografía: que-qué, porque.
Ortografía: segmentación.
Ortografía: ultracorreción.
Ortografía: verbos auxiliares.
Palabra incompleta.
Posible error: general
Semántica: error de sentido o régimen de una palabra.
Semántica: interferencias entre términos (cuanto/cuando menos).
Semántica: precisión de algunos términos que parecen sinónimos.
Semántica: precisión de palabras o expresiones.
Semántica: restricciones que impone una palabra sobre su cotexto o complementos (barajar, mayoría).
Términos casi homófonos (esotérico-exotérico).
Términos homófonos.
Tipografía.
Tipografía: Abreviaturas.
Tipografía: Horas.

FENÓMENOS
Tipografía: Mayúsculas.
Tipografía: Puntuación.
Tipografía: Siglas.

Anexo 12. Zonas geográficas

ZONA GEOGRÁFICA	
ID	**Localización**
1	Español panhispánico
2	Cuba
3	México
4	Chile, zona andina
5	Caribe
6	Argentina, zona rioplatense (zona de voseo)
7	Centroamérica
8	Costa Rica
9	Honduras
10	Centroamérica
11	Venezuela
12	Colombia
13	Panamá
14	R. Dominicana
15	Uruguay (zona de voseo)
16	Paraguay (zona de voseo)
101	Asturias
102	Andalucía
103	Cataluña
104	Centro de España
105	Este de España
106	Canarias
107	Zona noroccidental de España (León, Galicia, Asturias)
901	Algunas zonas de España
1001	Español de Europa
1002	Español de América

Anexo 13. Adverbios de grado

absolutamente
algo
aproximadamente
bastante
basto
casi
completamente
demasiado
enormemente
excepto
excesivamente
extremadamente
infinitamente
justamente
justo
más
mayormente
menos
mucho
muy
nada
poco
radicalmente
relativamente
salvo
sensiblemente
sobremanera
solamente
solo
suficiente
sumamente
tan
tanto
terriblemente
todo

Anexo 14. Adjetivos de grado extremo

abominable
atroz
brutal
colosal
delicioso
descomunal
divino
encantador
enorme
esencial
espantoso
espléndido
estupendo
excelente
excelso
eximio
exquisito
extraordinario
fabuloso
fantástico
fenomenal
formidable
fundamental
gélido
helado
horrible
horroroso
increíble
ínfimo
inmaculado
inmenso
insignificante
magnífico
maravilloso
máximo
mínimo
minúsculo
monstruoso
precioso
sensacional
supremo
terrible
tórrido
tremendo

Anexo 15. Locuciones latinas

a contrariis
a díe
a divinis
a fortiori
a látere
a nativitate
a pari
a posteriori
a priori
a quo
a símili
ab aeterno
ab initio
ab intestato
ab irato
ab ovo
ad bona
ad calendas graecas
ad cautélam
ad hoc
ad hóminem
ad honórem
ad líbitum
ad lítem
ad nútum
ad pédem lítterae
ad perpétuam
ad quem
ad referéndum
ad sénsum
ad valórem
álea jacta est
ante díem
ante merídiem
arate cavate
audaces fortuna juvat
áurea mediócritas
auris maris
consummátum est
córam pópulo
de jure
de verbo ad vérbum
de visu
de vita et móribus
Deo volente
deus ex máchina
do ut des
echar menos
ex abrupto
ex abundantia cordis
ex testamento
ferendae sententiae
honoris causa
id est
illícium vérum
in aetérnum
in ánima vili
in artículo mortis
in curia
in díem
in extenso
in extremis
in facie ecclesiae
in illo témpore
in íntegrum
in memóriam
in mente
in pártibus
in pártibus infidélium
in péctore
in perpétuum
in petto
in púribus
in saécula saeculórum
in sécula
in sécula seculórum
in situ
in statu quo
in utroque
in utroque jure
ínter nos
ínter vivos
interpósita persona
invita minerva

ipso facto
ipso jure
ite, missa est
latae sententiae
laus Deo
loco citato
magíster díxit
mane, thecel, phares
manu militari
mare mágnum
misti fori
mixti fori
modus operandi
mortis causa
mutatis mutandis
ne quid nimis
némine discrepante
níhil obstat
non sancta
non sancto
opera prima
pane lucrando
para in sécula
para sécula
peccata minuta
per áccidens
per ístam
per saécula saeculórum
per se
per sécula seculórum
Petrus in cunctis
plus minusve
plus ultra
prae mánibus
prima facie
pro domo súa
pro forma
pro indiviso
pro rata
pro rata parte
pro tribunali
própter nuptias
quid divínum
rara avis in terris
rata parte
relata réfero
requiéscat in pace
sine qua non
stábar máter
stábat máter
statu quo
sub júdice
sui géneris
súrsum corda
témpora, o mores
tiquis miquis
urbi et orbi
ut retro
ut supra
vade retro
vel cuasi
velis nolis
vera efigies
verbi gratia
vis cómica