Investigación cuantitativa de afijos y clíticos del español de México:

Glutinometría en el *Corpus del Español Mexicano Contemporáneo*

Alfonso Medina Urrea

Tesis para obtener el grado de
Doctor en Lingüística
Asesor: Luis Fernando Lara

Centro de Estudios Lingüísticos y Literarios
El Colegio de México, A. C.
México, D.F.
2003

Resumen

Esta tesis es, por un lado, una exploración cuantitativa del *Corpus del Español Mexicano Contemporáneo* y, por el otro, una investigación sobre la naturaleza de algunos conceptos lingüísticos observables en córpora mediante herramientas computacionales construidas especialmente para eso. Los objetivos fueron:

1. Determinar empírica y automáticamente los signos del nivel morfológico (los más posibles). Esto implica construir una serie de rutinas para examinar y comparar diferentes métodos de segmentación de palabras a partir de los cuales se pueda formalizar un procedimiento para la construcción de un catálogo de morfemas.

2. Construir un programa que segmente palabras a partir de los signos morfológicos del objetivo anterior (es decir, que separe las raíces de los afijos, *stemming*). En lo posible, hacer que el programa distinga entre las ocurrencias de un afijo y las ocurrencias, sin ser afijo, del segmento que lo representa (por ejemplo, que proponga 'aument-e' y no 'au-mente').

3. Seleccionar criterios para determinar los signos o palabras más gramaticales (aquellos más pertinentes a la estructura de la lengua).

4. Aplicar estos criterios para determinar dichos signos gramaticales (palabras función). Esto implica construir un programa que califique a cada palabra en términos de su cualidad, uso o funcionamiento gramatical.

5. Estudiar los métodos seleccionados arriba para examinar su posible integración o articulación a un esquema generalizador.

Estos objetivos se llevaron a cabo en gran medida: al lograr ordenar los signos del nivel morfológico —y, separadamente, los vocablos gráficos— según los valores de los índices propuestos como estimaciones de algunos rasgos característicos de los segmentos gramaticales, se cumplieron los objetivos 1 y 3. Los programas construidos para los objetivos 2 y 4 se basan en estos criterios e implican, necesariamente, un margen de error. Así, por ejemplo, se obtuvo un 95.5% de aciertos para el programa del objetivo 2. En cambio, para el segundo programa no hay evaluación. Los vocablos gráficos se ordenan en una lista encabezada por los segmentos más utilizados gramaticalmente, seguidos por aquellos que podrían considerarse de uso gramatical (que además alternan con segmentos de carácter gramatical dudoso) y finalmente por los que definitivamente no deben considerarse gramaticales porque son vocablos de contenido. La frontera no es clara, pero todo esto puede servir de partida para futuras caracterizaciones del concepto mismo de uso gramatical de un segmento. El último capítulo describe el esquema glutinométrico propuesto como generalización de los métodos utilizados en los primeros capítulos. En resumen, los logros de esta investigación se pueden enumerar:

- Se compararon diversos métodos de segmentación de palabras —estadística de digramas (prueba de independencia de χ^2, información mutua, razón de semejanza, coeficiente

de Yule), entropía, índices de cuadros y de economía de de Kock— y se mostró que estos últimos (aquellos que miden algún rasgo lingüístico de los afijos en general) son más confiables.

- A partir de estos métodos —que no son específicos para la lengua española—, se propuso y aplicó un índice formal apto para medir la cualidad de afijo (o *afijalidad*) de cualquier segmento de palabra o vocablo.

- Se extendieron estos métodos para medir el carácter de clítico (o *cliticidad*) de aquellos segmentos que, aunque gráficamente independientes (esto es, aparecen entre dos espacios), dependen formalmente de éstas. Se afinó el cálculo de esta propiedad al tomar en cuenta no uno sino dos valores de asociación (de la derecha y de la izquierda): al observar los datos, se hizo evidente que la cliticidad debe ser función de la diferencia de estos valores.

- Se formalizaron estos métodos o criterios. Es decir, se determinaron fórmulas de carácter matemático para extraer los datos morfológicos (relativos a la segmentación morfológica) y morfosintácticos (en relación a la asociación entre palabras y segmentos).

- Basándose en los criterios de *afijalidad*, se construyeron automáticamente diversos tipos de catálogos de signos afijales del *CEMC*.

- De manera similar, se seleccionaron los signos más gramaticales de ese corpus, es decir, aquellos más aptos de funcionar —según los criterios de *cliticidad* examinados— como elementos estructurales por su distribución en dicho corpus. Concretamente, así como la cliticidad verdadera es una función de dos valores de asociación (su resta), el carácter gramatical de cada segmento se puede estimar mediante la suma de dichos valores.

- Los catálogos de afijos, clíticos y las listas de palabras función obtenidas automáticamente a partir del *CEMC* constituyen un tipo de descripción formal de la lengua española, cuando menos de la hablada y escrita en México.

- La investigación de métodos para cuantificar tanto la afijalidad como la cliticidad de los segmentos de un corpus condujo a su generalización en la noción de *glutinosidad* o pegajosidad cuantitativa entre cadenas de morfemas al interior de la palabra y del sintagma. Se propuso un esquema glutinométrico (que especifica unidades de medición, bases axiomáticas, etc.) apto de medir dicha glutinosidad, cuando menos en lenguas como el español.

- Se construyeron diversos programas para los experimentos de descubrimiento y procesamiento de signos gramaticales tanto al interior como al exterior de la palabra gráfica, entre ellos un *tokenizer* o fichador (que filtra y segmenta el corpus, además de contabilizar la puntuación), un separador de raíces y afijos, un programa que ordena los vocablos gráficos según sus índices de entropía y economía y dos glutinómetros (uno para el interior y otro para el exterior de la palabra gráfica).

Índice general

Resumen i

Índices iii

Agradecimientos x

Lista de símbolos xi

Abreviaturas xiii

Introducción 1

0.1 Problema . . . 3
0.2 Objetivos . . . 9
0.3 Delimitación . . . 10
0.4 Metodología . . . 12
0.5 Premisas . . . 14
0.6 Hipótesis . . . 16
0.7 Plan de la tesis . . . 24

1 Panorama general 27

1.1 Análisis automático al exterior de la palabra . . . 27
1.2 Descubrimiento de vocablos gramaticales . . . 37
1.2.1 Las frecuencias de los vocablos . . . 38
1.2.2 Otros métodos de adquisición léxica . . . 42
1.3 Morfología automática . . . 48
1.3.1 Fonología de estados finitos . . . 52
1.3.2 Medidas estadísticas de productividad de reglas morfológicas . . . 56
1.4 Segmentación automática de palabras . . . 61
1.4.1 Reconocimiento de patrones . . . 62
1.4.2 Métodos de estadística de digramas . . . 72
1.4.3 Frecuencias de caracteres (la escuela rusa) . . . 75
1.4.4 Cuentas de fonemas anteriores y posteriores . . . 80
1.4.5 Teoría de la información . . . 83
1.4.6 El principio de economía . . . 87
1.5 Observaciones finales . . . 90

2 El afijo en el *CEMC* **91**
2.1 Sobre las unidades morfológicas 92
2.2 Cuestiones preliminares 96
2.2.1 Nociones formales preliminares 97
2.2.2 Rutinas y estructuras de datos 99
2.3 Índices para cuantificar la *afijalidad* de una segmentación 103
2.3.1 Número de cuadros 104
2.3.2 Entropía 106
2.3.3 Índice de economía 110
2.3.4 Medidas estadísticas 116
2.4 Comparación de índices 119
2.5 El catálogo de afijos 122
2.5.1 Definición formal 123
2.5.2 Probabilidades 124
2.6 Hacia un índice de *afijalidad* 126
2.7 Catálogos de afijos a partir del *CEMC* 130
2.8 Los sufijos del español de México 143
2.9 Observaciones finales 169

3 El clítico en el *CEMC* **172**
3.1 Sobre los signos gramaticales 173
3.2 El clítico como pariente del afijo 177
3.2.1 Número de cuadros 179
3.2.2 Entropía 181
3.2.3 Índice de economía 183
3.3 Un índice de puntuación 185
3.4 Hacia un índice de *cliticidad* 188
3.5 Cuestiones preliminares en la determinación de los clíticos del *CEMC* 192
3.5.1 Las cadenas de Markov 192
3.5.2 Procedimiento y discusión 194
3.6 Resultados del procedimiento 199
3.7 Los clíticos en el *CEMC* 220
3.8 Observaciones finales 231

4 Glutinometría en el *CEMC* **234**
4.1 Antecedentes 235
4.2 La lógica del esquema de *glutinosidad*: hacia un índice cuantitativo 241
4.2.1 La afijalidad y la cliticidad como fuerzas opuestas 242
4.2.2 La afijalidad y la cliticidad como dos instancias de la misma fuerza . . 245
4.2.3 Las fronteras del sintagma y el carácter infinito del lenguaje 246
4.2.4 Discurso y diacronía 250
4.3 Teoría de la medición: hacia una glutinometría formal 252
4.3.1 Generalidades de la medición 253
4.3.2 Las magnitudes y las dimensiones de la medición 261
4.3.3 Las unidades de medición 278

4.3.4 Las bases axiomáticas de la medición . . . 285
4.3.5 El problema del error . . . 294
4.4 Glutinometría al interior del sintagma en el *CEMC* . . . 301
4.5 Los signos gramaticales del español de México . . . 311
4.6 Observaciones finales . . . 331

Conclusiones **334**
4.6.1 Sinopsis de experimentos . . . 336
4.6.2 Las hipótesis de glutinosidad reformuladas . . . 339
4.6.3 Problemas, ventajas y pendientes de una glutinometría . . . 340
4.6.4 Conclusiones . . . 346

A El *Corpus del Español Mexicano Contemporáneo* **350**
A.1 Descripción . . . 350
A.2 Preprocesamiento . . . 352
A.3 Las marcas gramaticales del *CEMC* . . . 358
A.4 Formas más frecuentes . . . 358
A.4.1 Vocablos . . . 360
A.4.2 Abreviaturas . . . 360

B Muestra aletoria de vocablos analizados **374**

C Sufijos en el *CEMC* **398**

D Las formas más gramaticales del *CEMC* **421**

Bibliografía **436**

Índice de materias **446**

Índice de tablas

1.1 Análisis estadístico de los vocablos del *CEMC* 39
1.2 Las formas sin lematizar del *CEMC* con frecuencias absolutas mayores 41
1.3 Digramas más frecuentes en un corpus del *New York Times* 45
1.4 Productividad *strictu sensu* y utilidad pragmática de reglas morfológicas en la formación de sustantivos holandeses 57
1.5 Segmentaciones posibles del vocablo 'zerlegen' 69
1.6 Algunas reglas del algoritmo de Porter 71
1.7 Tabla de contingencia para el digrama $w_1 w_2$. 73
1.8 Estadísticas para medir no asociación entre digramas. 74
1.9 Frecuencias de caracteres de la biblia en alemán 78
1.10 Función correlativa *KF* de los caracteres más frecuentes según su posición en la cadena de caracteres 78
1.11 Cuentas de fonemas anteriores y posteriores en cada segmentación de *What did he think of?* 82
1.12 Hipótesis para cada corte de los vocablos 'capacidad' y 'olvidad' 89

2.1 Tipos de combinaciones de segmentos 105
2.2 Entropía de la segmentación $p::B_{i,1}$ 108
2.3 Valores de entropía en cada segmentación del vocablo 'aparecer'. 109
2.4 Tabla de contingencia para el digrama 'previa::mente'. 116
2.5 Medidas estadísticas de cada segmentación de 'previamente'. 118
2.6 Medidas estadísticas de cada segmentación de 'aparecer'. 119
2.7 Comparación de índices: segmentaciones correctas en una muestra de 836 vocablos. 120
2.8 Medidas de segmentación del vocablo 'aumente'. 132
2.9 Medidas de segmentación del vocablo 'comente'. 132
2.10 Medidas de segmentación del vocablo 'previamente'. 132
2.11 Medidas de segmentación del vocablo 'nacionalidad'. 133
2.12 Selección de sufijos del español según el *CEMC* en orden de *afijalidad* 138
2.13 Selección de prefijos del español según el *CEMC* en orden de *afijalidad* 140
2.14 Sufijos de flexión nominal 145
2.15 Sufijos de flexión verbal del modo indicativo 146
2.16 Flexiones del subjuntivo 149
2.17 Sufijos de verboides 150
2.18 Los verbos y derivación (con y sin marcas de flexión) 151
2.19 Grupos de sufijos con marca adverbial 152

2.20 Grupos de sufijos derivativos nominales (según parecido formal) 154
2.21 Enclíticos descubiertos como sufijos gráficos 165
2.22 Gerundio y enclíticos 166
2.23 Imperativo y enclíticos 166
2.24 Infinitivo y enclíticos 167

3.1 Pesos asignados a los signos de puntuación 188
3.2 Entropía que cabe esperar después del segmento 'de' 197
3.3 Preformas gramaticales del *CEMC* en orden de *cliticidad* 201
3.4 Postformas gramaticales del *CEMC* en orden de *cliticidad* 204
3.5 Proclíticos del *CEMC* en orden de *cliticidad* 207
3.6 "Enclíticos" del *CEMC* en orden de *cliticidad* 210
3.7 Formas gramaticales del *CEMC* en orden de *cliticidad* total 213
3.8 "Nexos" del *CEMC* 216
3.9 Las 30 formas más "enclíticas" del *CEMC* 222
3.10 Las 30 formas más proclíticas del *CEMC* 228

4.1 El mesurando y sus estimaciones 255
4.2 Correspondencia entre los grados de una propiedad y sus equivalentes instrumentales con respecto a los números 258
4.3 Relación entre cuantificación y medición con respecto a los números 259
4.4 Dimensiones de la Glutinometría 263
4.5 Selección de sufijos con más entropía que economía 277
4.6 Selección de sufijos con más economía que entropía 278
4.7 Unidades de la Glutinometría 285
4.8 Formas gramaticales del *CEMC* 302
4.9 Proclíticos del *CEMC* 305
4.10 Formas del *CEMC* con cliticidades cercanas 308
4.11 Promedios de rangos por categorías gramaticales 314
4.12 Esquema simplificado de paradigmas de formas gramaticales (el sintagma nominal) 315
4.13 Pronombres personales 319
4.14 Esquema simplificado de paradigmas de formas gramaticales (la oración) . . . 320
4.15 Tipos de adverbios 321
4.16 Rasgos de algunos adverbios 322
4.17 Oposiciones entre adverbios 323
4.18 Los verbos y sus formas flexionadas 324
4.19 Nociones espacio-temporales 326
4.20 Objetos concretos y abstractos 327
4.21 Propiedades 328
4.22 Formas léxicas 329

A.1 Géneros en el *CEMC* 351
A.2 Caracteres con varias funciones en el *CEMC* 353
A.3 Caracteres menos ambiguos en el *CEMC* 354
A.4 Aspecto de un fragmento del *CEMC* 355

A.5 Aspecto de un fragmento del archivo CEMC2.TXT 356
A.6 Frecuencias y porcentajes de caracteres en CEMC2.TXT 357
A.7 Modificaciones a caracteres para reflejar correspondencia entre grafemas y fonemas 358
A.8 Marcas gramaticales del *CEMC* 359
A.9 Los lemas más frecuentes del *CEMC* (frecuencias corregidas) 362
A.10 Las formas más frecuentes en CEMC2.TXT 365
A.11 Abreviaturas en el *CEMC* 370

B.1 Vocablos de la muestra omitidos 375
B.2 Resultados del análisis de la muestra 375
B.3 Correspondencias de signos para la tabla B.4 377
B.4 Muestra aletoria de vocablos analizados 379

C.1 Sufijos del *CEMC* en orden de *afijalidad* 400
C.2 Algunas agrupaciones por forma de sufijos derivativos observadas en la tabla C.1 que no se examinaron en el capítulo sobre el afijo. 419

D.1 Formas gramaticales del *CEMC* 423

Índice de figuras

1.1 Transductor de estados finitos para la regla t:c ⇒ _i:i 53
1.2 Gráfica de la entropía de dos mensajes posibles $X = \{x_1, x_2\}$ 85
1.3 Esquema para ilustrar las probabilidades de los segmentos que según un corpus ocurren después de *elabor~*. 86

2.1 Representación de las segmentaciones posibles de un vocablo x (v_x). 98
2.2 Combinaciones de segmentos de la izquierda y de la derecha 112
2.3 Distribución de los valores de afijalidad (sufijalidad) de todos los segmentos recogidos en el catálogo de sufijos del español de México. 136
2.4 Distribución de los valores de afijalidad (prefijalidad) de todos los segmentos recogidos en el catálogo de prefijos del español de México. 137

3.1 Anidamiento de signos al interior y al exterior de la palabra 178
3.2 Cadena de Markov de primer orden para calcular entropías de cada vocablo . . 182
3.3 Cadena de Markov para calcular entropías *en reversa* 183
3.4 Cadena de Markov arbórea (de varios órdenes) 184
3.5 Cadena de Markov con información cuantitativa de los signos de puntuación . 186
3.6 Esquema arbóreo de frase preposicional 223

4.1 Fronteras entre morfemas en la oración *Dogs were indisputably quicker.* 239
4.2 Glutinosidad más alta al interior de la palabra 243
4.3 Afijalidad y cliticidad como instancias de la misma fuerza 246
4.4 La glutinosidad en el plano cartesiano 293

Agradecimientos

Esta tesis se llevó a cabo gracias al patrocinio parcial de El Colegio de México A.C., el Sistema Nacional de Investigadores (CONACyT), el Servicio Alemán de Intercambio Académico (Deutscher Akademischer Austauschdienst, DAAD) y el Grupo de Ingeniería Lingüística del Instituto de Ingeniería (GIL-IINGEN) de la Universidad Nacional Autónoma de México. A El Colegio de México le debo además el permiso de utilizar el corpus completo del español mexicano contemporáneo (*CEMC*).

Quiero agradecer en especial la orientación y ayuda de mi asesor el Dr. Luis Fernando Lara, así como el apoyo siempre oportuno de la Dra. Martha Elena Venier. También quiero expresar mi gratitud a los Doctores María Pozzi Pardo, Marianna Pool Westgaard y Pedro Martín Butragueño por la lectura crítica y cuidadosa de este trabajo. Agradezco, además, a las Doctoras Concepción Company Company, Josefina García Fajardo y María Eugenia Vázquez Laslop por sus valiosas observaciones sobre cuestiones aisladas relacionadas con este trabajo. Asimismo, estoy en deuda con Isabel García Hidalgo, Ida Courtade Bevilacqua, Elsa Cristina Buenrostro Díaz y María del Carmen Larios Lozano por su aliento y comentarios también sobre cuestiones aisladas.

El patrocinio del DAAD me permitió trabajar en la Universidad de Tréveris (Fach II Linguistische Datenverarbeitung), donde encontré la orientación y el apoyo del Profesor Dr. Burghard Rieger. Agradezco también la oportuna asistencia del Profesor Dr. Reinhard Köhler. Tampoco olvido el auxilio que me brindaron en diferentes momentos Christiane Hoffmann y los Doctores Sven Naumann y Alexander Mehler.

Asimismo, gracias al Dr. Gerardo Sierra Martínez, director del Grupo de Ingeniería Lingüística del Instituto de Ingeniería e investigador responsable del proyecto *Desarrollo del Corpus Lingüístico en Ingeniería* (CONACyT R37712-A) cuyo patrocinio me permitió realizar las últimas correcciones de las páginas siguientes, contribución decisiva a la conclusión de este trabajo.

Por último, agradezco el apoyo crucial y generoso de la Dr. Rebeca Barriga Villanueva, así como el de Blanca y María Teresa Medina Carbajal, Alejandro Reza Arriaga y Alexandre Toshirrico Cardoso Taketa.

Lista de símbolos

Símbolo	*Significado*
Ψ	secuencia de ocurrencias de palabras (corpus)
$\xi = \|\Psi\|$	número de ocurrencias de palabras en Ψ
V	conjunto de vocablos
F	conjunto de frecuencias
Φ	conjunto de vocablos y sus frecuencias
$v_i \in V$	vocablo i del conjunto V
$f(v_i) = f_i \in F$	frecuencia de vocablo v_i miembro del conjunto F
s_k	vocablo k o fragmento k de un vocablo x
$f(s_k)$	frecuencia del segmento s_k
$f(\overline{s_k})$	frecuencia de los segmentos que no son s_k
$\langle v_i, f_i \rangle \in \Phi$	par ordenado miembro de Φ
$a_{i,j}::b_{i,j}$	segmentación j de vocablo v_i
$A_{i,j}$	conjunto de segmentos de la izquierda que combinados con segmento $b_{i,j}$ forman vocablos presentes en V
$B_{i,j}$	conjunto de segmentos de la derecha que combinados con segmento $a_{i,j}$ forman vocablos presentes en V
$A^p_{i,j}$	conjunto de segmentos de la izquierda que son prefijos hipotéticos del segmento $b_{i,j}$
$B^s_{i,j}$	conjunto de segmentos de la izquierda que son sufijos hipotéticos del segmento $a_{i,j}$
$c_{i,j}$	número de cuadros en la segmentación j del vocablo v_i
$k^p_{i,j}$	medida de la economía de la segmentación j del vocablo v_i cuando el segmento de la izquierda se supone prefijo
$k^s_{i,j}$	medida de la economía de la segmentación j del vocablo v_i cuando el segmento de la izquierda se supone sufijo
$h^p_{i,j}$	entropía de prefijo en segmentación j de vocablo v_i
$h^s_{i,j}$	entropía de sufijo en segmentación j de vocablo v_i
Υ^p	catálogo de prefijos
Υ^s	catálogo de sufijos
$\gamma = \|\Upsilon\|$	número de afijos en catálogo Υ
$\Omega = \|V\|$	número de vocablos en V
Ω'^p_k	número de vocablos que empiezan con segmento s_k
Ω''^p_k	número de vocablos que empiezan con segmento s_k en calidad de prefijo

Símbolo	*Significado*
$\bar{c}_k$	promedio de cuentas de cuadros asociados al afijo k
$\bar{k}_k$	promedio de índices de economía asociados al afijo k
$\bar{h}_k$	promedio de índices de entropía asociados al afijo k
AF	afijalidad
CL	cliticidad
GL	glutinosidad
r_x	resistencia a la glutinosidad causada por la puntuación probable antes o después de un segmento s_x
α_i	objeto lingüístico de la lengua del corpus Ψ
$\mathbf{nuevos}_x$	número de signos nuevos de un nivel formado por la combinación del signo s_x (del nivel inferior) con otros signos (también del nivel inferior)
$\mathbf{alternantes}_x$	tamaño del conjunto de signos que alternan con el signo s_x (están en distribución complementaria) y que se combinan con otros signos para formar signos nuevos del siguiente nivel
I	dimensión de la entropía, medida en *bits*
S	dimensión del número de signos, medida en signos de varios tipos
G	dimensión a la que pertenecen las magnitudes de afijalidad, cliticidad y glutinosidad
Kock	unidad que mide la economía de una segmentación
Varrón	unidad que mide la glutinosidad

Abreviaturas

abrev.	abreviatura
adj.	adjetivo
adv.	adverbio
afjdad.	afijalidad
art.	artículo
c.	cantidad
car.	caracter
cdrs.	cuadros
CELL	Centro de Estudios Lingüísticos y Literarios (COLMEX)
CEMC	*Corpus del Español Mexicano Contemporáneo*
coef.	coeficiente
COLMEX	El Colegio de México, A. C.
conj.	conjunción
copret.	copretérito
cuant.	cuantificador
DEM	Diccionario del Español de México
det.	determinativo
der.	derecha
ej.	ejemplo
econ.	economía
encl.	enclítico
entrop.	entropía
fr.	frecuencia
i.m.	información mutua
imp.	imperativo
inf.	infinitivo
info.	información
izq.	izquierda
lat.	latín
n.	nombre
núm.	número
o	oración
p.	persona (ej. 3ªp.)
pl.	plural
prep.	preposición

pret.	pretérito
prob.	probabilidad
pron.	pronombre
punt.	puntuación
r.s.	razón de semejanza
rel.	relativo
sing.	singular
SN	sintagma nominal
SP	sintagma preposicional
SV	sintagma verbal
subj.	subjuntivo
subs.	substracción
subord.	subordinador
sust.	sustantivo
v.	verbo

Introducción

> Every linguist knows what linguistics is; he knows all about the aims of his own research and usually he also knows what other linguists are doing. He hopes that if he does what others do, he will be performing science. However, he seldom thinks about what science is, what its components are and what the place of linguistics in the system of science is. **Gabriel Altmann**

Hace más de veinte años que el primer analizador sintáctico automático de la lengua española vio la luz en El Colegio de México. El proyecto —ambicioso para la época, dados sus objetivos y los recursos entonces disponibles— auxilió en el análisis gramatical del *Corpus del Español Mexicano Contemporáneo* (esto es, en la asignación de marcas gramaticales a cada palabra)[1]. De entonces a la fecha, el campo del análisis automático ha crecido considerablemente, especialmente en el nivel sintáctico y en lo que se refiere a métodos basados en la introspección o en información disponible a priori sobre las lenguas privilegiadas (generalmente indoeuropeas y predominantemente la inglesa).

Pero la disponibilidad de una colección de textos como el *Corpus del Español Mexicano Contemporáneo* (de aquí en adelante *CEMC*), en cuya compilación se tomaron en cuenta criterios que aseguraran cierto grado de representatividad[2], y el progreso constante de las

[1] Véanse Luis Fernando Lara, Roberto Ham Chande y Ma. Isabel García Hidalgo, eds., *Investigaciones lingüísticas en lexicografía* [89], El Colegio de México, México, (*Jornadas* 89) 1979 y Luis Fernando Lara, *Dimensiones de la lexicografía. A propósito del Diccionario del Español de México* [85], El Colegio de México, México, (*Jornadas* 116) 1990.

[2] Sobre la representatividad del *CEMC* en concreto, véanse Lara y Ham, "Base estadística del DEM" [88],

nuevas tecnologías, que hacen posible el procesamiento de córpora considerablemente más grandes, son una clara invitación a investigar métodos automáticos de extracción de datos cuantitativos de la lengua española a partir de la menor información a priori posible: es decir, tomando en cuenta los rasgos que son característicos no sólo del español o de sus parientes cercanos, sino de la mayoría de las lenguas; aquellos datos que describen más lo universal y menos lo peculiar de esta lengua. Por ejemplo, en lugar de partir del *conocimiento* particular y tradicional de los afijos y clíticos del español, en este trabajo se investiga (mediante criterios cuantitativos, de ninguna manera específicos a esta lengua) el comportamiento de los segmentos que funcionan como afijos o clíticos de otros segmentos.

Esta tesis es —por un lado— un intento de reunir *automáticamente* a partir del *CEMC* los datos lingüísticos mínimamente necesarios (sobre todo morfológicos, pero también de carácter morfosintáctico) para construir una estructura formal descriptiva del español, para lo cual —por el otro— se trata también de una investigación sobre la naturaleza cuantitativa de algunos conceptos lingüísticos aplicables al lenguaje en general.

En esta introducción se esboza el trabajo presentado en los capítulos siguientes. A continuación se describen las líneas generales de la tesis: el planteamiento del problema, objetivos, delimitación, premisas, hipótesis y metodología del proyecto. La última sección de esta introducción es un resumen de lo hecho en cada capítulo.

---

pp. 5-39 de Lara, Ham y García, *op. cit.* [89] 1979, y Lara, "Caracterización metódica del *corpus* del DEM" [86], pp. 85-106 de Lara, *op. cit.* [85] 1990; sobre los conceptos de ***representatividad*** y ***corpus balanceado*** véase Manning y Shütze, ***Foundations of Statistical Natural Language Processing*** [93], The MIT Press, Cambridge (Mass.), 1999, pp. 19, 119-120.

0.1 Problema

Construir un analizador automático, ya sea morfológico o sintáctico, para una lengua natural no es en sí un problema de investigación. Ya existen numerosos sistemas basados en distintos métodos que, con mayor o menor éxito, llevan a cabo diferentes tipos de análisis automático[3]. El verdadero problema es más bien cómo y de dónde viene la descripción del fenómeno en que se base dicho análisis.

Este problema consta de varios aspectos: primero, uno empírico[4] de extraer a partir de un corpus las unidades lingüísticas pertinentes; segundo, el aspecto conceptual de definir esas unidades para poder extraerlas y utilizarlas; tercero, el problema metodológico de escoger las estrategias convenientes para llevar a cabo lo anterior; y, finalmente, el problema de determinar y evaluar los criterios que nos permitan calificar lo apropiado que pueda tener un dato al clasificarlo como caso de tal o cual concepto lingüístico.

Por lo general, la mayoría de los métodos de análisis automático sintáctico resuelven los dos primeros aspectos del problema presuponiendo que lo que se sabe de la gramática es definitivo; es decir, que la gramática ya es algo dado, que las unidades lingüísticas son conocidas y transparentes y que la introspección es un método "empírico"[5] y suficiente para la investigación del lenguaje. Así, los métodos más favorecidos dependen de reglas que sim-

[3]Por ejemplo, véase la diversidad de métodos en manuales como los de Allen (*Natural Language Understanding* [4], Benjamin/Cummings, Redwood, 1995); y específicamente de métodos cualitativos: Naumann y Langer (*Parsing: Eine Einführung in die maschinelle Analyse natürlicher Sprache* [108], Teubner, Stuttgart, 1994), Gazdar y Mellish (*Natural Language Processing in Prolog* [52], Addison-Wesley, Wokingham, Gran Bretaña, 1989); y cuantitativos: Charniak (*Statistical Language Learning* [32], The MIT Press, Cambridge (Mass.), 1993).

[4]En el sentido de *factual*: "que se refiere a estados de hecho"; véase Nicola Abbagnano, *Diccionario de filosofía* [1], tr. Alfredo N. Galletti, Fondo de Cultura Económica, México, 1991 [1961], *s.v.*

[5]Pero en el sentido de experiencia *intuitiva*; véase *ibid.* [1], *s.v.*, EMPÍRICO.

bolizan la concepción que el analista tiene de la lengua en cuestión. Esto quiere decir que se consultan las gramáticas tradicionales disponibles o, lo más frecuente, que se recurre a la introspección para construir dichas reglas[6], cosa en mi opinión poco científica, no porque el método introspectivo no sea valioso para la ciencia, sino porque al depender casi exclusivamente de éste se desatiende buena parte de los hechos reales. Por otra parte, los proyectos más empíricos —aquellos construidos a partir de hechos lingüísticos (datos documentados)— requieren casi siempre de córpora ya marcados, es decir, con anotaciones gramaticales previamente aplicadas manualmente a cada una de sus palabras[7]. Sobra señalar que dichos recursos son de disponibilidad sumamente limitada.

Otra cuestión que incumbe al trabajo empírico inherente al análisis automático es la selección del nivel lingüístico que se quiere investigar. Por ejemplo, un nivel muy a menudo descuidado es el morfológico. De hecho, cuando siquiera se considera, la tendencia general es, como en el nivel sintáctico, presuponer la existencia de los morfemas (darlos por bien conocidos) antes de empezar la investigación. Esto se explica en parte porque el inglés, la lengua más trabajada en el campo de la automatización, tiene una morfología muy sencilla y muy conocida[8], cosa que también explica el poco interés en los métodos automáticos de descu-

---

[6]En gran medida, el campo del análisis automático heredó de la lingüística generativa la idea de que los juicios e intuiciones de un hablante nativo de una lengua (derivados de un proceso de introspección, es decir, basados en la gramática mental internalizada de ese hablante) son suficientes para el quehacer del lingüista. Así, aunque a veces se recurre a la consulta de las gramáticas tradicionales de la lengua estudiada, la mayoría de los trabajos de corte computacional se basan en la formulación de reglas gramaticales a partir de los juicios e intuiciones de quienes las formulan.

[7]Lo que en inglés se conoce como entrenamiento supervisado del sistema (*supervised training*) y que se refiere al hecho de proporcionarle manualmente al programa la información gramatical de los datos de una muestra. Cuando la clasificación de dichos datos no está disponible, se habla de aprendizaje o entrenamiento no supervisado (*unsupervised training*). Este trabajo se orienta hacia la segunda estrategia de investigación. Véase Manning y Schütze, *op. cit.* [93] 1999, pp. 232.

[8]De allí la popularidad de esquemas tan simples como el de Porter ("An Algorithm for Suffix Stripping" [115], *Program*, 14:3 (1980), pp. 130-137; Frakes y Baeza, "Stemming Algorithms" [49] en *Information*

brimiento de unidades morfológicas en general[9]. Curiosamente, muchos trabajos de análisis automático suelen preferir (y presuponer) cuestiones más abstractas, menos sistemáticas y más pertinentes al fenómeno de la significación que al gramatical (diversos temas típicos de la semántica e incluso reflexiones con aspiraciones presuntamente cognoscitivas). Esto es paradójico porque todas estas cuestiones dependen en alguna medida de la morfología de la lengua (en unas más que en otras; en español, por ejemplo, la dependencia es ciertamente significativa). De allí la necesidad de por lo menos estudiar los métodos para determinar morfemas automáticamente a partir de un corpus.

Aparte del aspecto empírico, y muy en relación con la selección del nivel lingüístico apropiado, está el problema conceptual de definir lo que se quiere extraer del corpus. Para un trabajo de procesamiento cuantitativo del nivel morfológico que suponga la mínima intervención del analista parece natural optar por los afijos y por los clíticos. Aunque estos conceptos se definirán apropiadamente más adelante (en los capítulos correspondientes), podemos adelantar definiciones provisionales: primero, entenderemos por 'afijo' aquel fragmento de palabras

---

Retrieval: Data Structures and Algorithms, Prentice Hall, New Jersey, 1992, pp. 131-160) que examinaremos adelante (a partir de la página 70). A pesar de sus deficiencias, el algoritmo de Porter funciona relativamente bien para la lengua inglesa. Sin embargo, las ***traducciones*** a otras lenguas como el español dejan mucho que desear.

[9]Aunque existen métodos de segmentación automática de palabras que veremos con detalle en el capítulo sobre el afijo —véanse por ejemplo en la bibliografía los trabajos de Hafer y Weiss ("Word Segmentation by Letter Successor Varieties" [60], *Information Storage and Retrieval*, 10 (1974), pp. 371-385, basado en el de Zellig Harris, "From Phoneme to Morpheme" [65], *Language* 31:2 (1955), pp. 190-222), Oliver Cromm (sobre el método de N. D. Andreev: ***Affixerkennung in deutschen Wortformen. Eine Untersuchung zum nicht-lexikalischen Segmentierungsverfahren von N. D. Andreev***, Abschluß des Ergänzungsstudiums Linguistische Datenverarbeitung [46], Francfort del Meno, 1996), Josse de Kock, y W. Bossaert (***Introducción a la lingüística automática en las lenguas románicas*** [80], Gredos, Madrid, 1974; ***The Morpheme. An Experiment in Quantitative and Computational Linguistics*** [82], Van Gorcum, Amsterdam/Madrid, 1978), Ursula Klenk, ed. (***Computation Linguae I, II***, [75, 77], Steiner, Stuttgart, 1992 y 1994), y Kyo Kageura ("Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences" [74], ***Journal of Quantitative Linguistics*** 6(1999), pp. 149-166)—, estos son relativamente pocos y no constituyen un problema resuelto mientras no se comparen para determinar sus diferencias y ventajas.

o vocablos[10] que es morfológicamente pertinente y que ocurre adherido a la base, ya sea al principio ('prefijo') o final ('sufijo') de las palabras o vocablos que forman; y, segundo, con el término 'clítico' se designará a las partículas que no constituyen palabras propiamente plenas y que, sin ser afijos, tienden a acompañar subordinadamente a otras palabras o vocablos.

Esto está íntimamente ligado con la distinción clásica entre palabras léxicas o de contenido (plenas) y palabras función o gramaticales (no plenas). Como se verá a lo largo de este trabajo, hay un grupo considerable de cosas que no encajan perfectamente entre estos dos tipos ideales de palabras. En su calidad de fragmentos del corpus, me referiré a estos tipos como 'segmentos' de contenido (palabras léxicas o plenas) y 'segmentos' gramaticales (palabras función). En el mismo sentido, todo lo que no es ni de contenido ni gramatical, o es ambas cosas, puede llamarse simplemente 'segmento'. Similarmente, me referiré tanto a bases y afijos con el mismo término, ya que además de segmentos del corpus, son segmentos de palabras.

Otra parte del problema es cómo interpretar, para determinar los signos gramaticales, lo que representa un corpus lingüístico, es decir, como concebimos la estructura lingüística implícita en él. Obviamente se trata de una cadena de signos, un fragmento de habla infinita, una muestra de lo que los hablantes suelen escuchar o leer. Pero esa simple secuencia de signos tiene implícita una estructura gramatical compleja que permite comunicar (o no) significados concretos y abstractos, particulares y generales, comunes y extraordinarios, y —especialmente importante para este trabajo— significados estructuradores del discurso (o gramaticales) y

[10]Como veremos adelante, el término 'palabra' se referirá al conjunto de letras (o fonemas) que ocurren en la cadena hablada y el término 'vocablo' se referirá al conjunto de ocurrencias de las palabras en esa cadena. La formalización de esta distinción se encuentra a partir de la página 97.

aquellos reveladores del contenido particular del discurso que estructuran.

De esta manera, cuando oímos hablar a alguien, puede ser que escuchemos cosas que nos sorprenden, que nos intrigan, que nos informan o, en su defecto, que simplemente ya sabemos. Similarmente, cuando leemos un texto nos encontramos con cadenas de caracteres que nos desconcertan por inesperadas, que nos intrigan porque no las entendemos, porque no las conocemos o porque ocurren en lugares inesperados.

Pero muchas de las cosas que oímos o leemos no nos causan asombro ni nos intrigan porque ya sabíamos que iban a ser dichas, es decir, las estábamos esperando (tarde o temprano tenían que ocurrir). Y una gran parte de ellas constituye, de hecho, una especie de vehículo gramatical que transporta el mensaje del discurso.

Obviamente son los segmentos gramaticales (artículos, preposiciones, marcas de flexión y derivación, etc.) los que constituyen este vehículo que facilita la comunicación de intrigas, desdichas, placeres o simple información que pueda depararnos algún texto hablado o escrito. En otras palabras, los segmentos gramaticales proporcionan una estructura que matiza a los de contenido (que son típicamente sustantivos y verbos).

Todo esto apunta hacia cierta oscilación entre lo común y lo inesperado en el discurso, que debe estar retratada en el corpus y que puede medirse cuantitativamente. De hecho, como veremos adelante, se ha mostrado repetidamente con otras lenguas que al medir la impredecibilidad de ocurrencia de ciertos segmentos en ciertos contextos se *descubren* fronteras morfológicas[11].

[11] Véanse por ejemplo: Hafer y Weiss, art. cit. [60] 1974; Frakes, art. cit. [49] 1992; el reporte que hace Michael Oakes del trabajo de Joula, Hall and Boggs 1994 (en *Statistics for Corpus Linguistics* [111], Edinburgh University Press, 1998, pp. 86-87), etc.

Sin embargo, sobra decir que esta oscilación entre lo esperado y lo incierto no es exclusiva del discurso propiamente lingüístico: hay comunicación sin palabras. Es más, también hay palabras sin comunicación (o con poca), esto es, estructuras lingüísticas sin significado propiamente informativo. De allí la importancia de tomar en cuenta otros aspectos de la estructura lingüística implícita en un corpus, para determinar las mejores maneras de segmentar palabras morfológicamente, como la capacidad combinatoria de los signos: si bien los segmentos gramaticales dicen por sí mismos poco en cuanto al contenido del discurso se refiere, los segmentos de contenido dicen mucho más al combinarse con los primeros. De hecho, un sinfín de segmentos léxicos podrá aglutinarse con algunos pocos segmentos gramaticales para convertirse en una infinidad mayor de signos de contenido de otro nivel. Es obvio que no cualquier segmento de contenido se combina con cualquier segmento gramatical: los patrones combinatorios son estructuras relativamente rígidas particulares a cada lengua en complejidad y flexibilidad. De todos modos, la aglutinación de signos de contenido con signos gramaticales resulta en estructuras que hacen de los sistemas lingüísticos, sistemas sumamente económicos.

Por supuesto, esto no agota la complejidad de lo que podemos encontrar en un corpus. Pero si en el corpus hay algo más que una secuencia intermitente de palabras gráficas, vale la pena averiguar si los datos cuantitativos que se pueden obtener de allí las reflejan y si en efecto constituyen pistas importantes que nos permitan descubrir afijos y clíticos automáticamente a partir de un corpus.

0.2 Objetivos

A continuación se hacen explícitos los objetivos de la tesis. Como se dijo arriba, los generales son, primero, reunir automáticamente a partir del corpus los datos lingüísticos mínimamente necesarios para construir un tipo de descripción de una lengua (a partir de afijos y clíticos) y, segundo, investigar las herramientas cuantitativas pertinentes a esa descripción. A estos se les pueden agregar los siguientes objetivos que los caracterizan y complementan y que exhiben el desarrollo progresivo de la investigación, partiendo del nivel morfológico, hacia la generalización de herramientas conceptuales:

1. Determinar empírica y automáticamente los signos del nivel morfológico (los más posibles). Esto implica construir una serie de rutinas para examinar y comparar diferentes métodos de segmentación de palabras a partir de los cuales se pueda formalizar un procedimiento para la construcción de un catálogo de morfemas.

2. Construir un programa que segmente palabras a partir de los signos morfológicos del objetivo anterior (es decir, que separe las raíces de los afijos, *stemming*). En lo posible, hacer que el programa distinga entre las ocurrencias de un afijo y las ocurrencias, sin ser afijo, del segmento que lo representa (por ejemplo, que proponga 'aument-e' y no 'au-mente').

3. Seleccionar criterios para determinar los signos o palabras más gramaticales (aquellos más pertinentes a la estructura de la lengua).

4. Aplicar estos criterios para determinar dichos signos gramaticales (palabras función). Esto implica construir un programa que califique a cada palabra en términos de su cualidad, uso o funcionamiento gramatical.

5. Estudiar los métodos seleccionados arriba para examinar su posible integración o articulación a un esquema generalizador.

Los primeros objetivos constituyen una investigación cuantitativa y empírica del corpus que involucra la construcción de diversas herramientas computacionales. El primero y el segundo son los objetivos del capítulo sobre el afijo. El tercero y el cuarto son los de aquel

sobre el clítico. El último corresponde a una reflexión integradora de los métodos a ser estudiados y es el objetivo del último capítulo de la tesis.

0.3 Delimitación

Después de definir las metas específicas de la investigación, la diversidad de enfoques posibles para alcanzarlas hace necesario delimitar el alcance de dichos objetivos. Por ejemplo, es importante enfatizar que el interés primordial es lingüístico, cosa que no necesariamente simplifica el proyecto al eliminar del centro de atención cuestiones de carácter ingenieril (tales como la eficiencia de un programa en términos de tiempo y espacio[12]). Esto es, el problema no se hace más sencillo, porque, al enfocar el fenómeno del lenguaje, se hace más evidente su naturaleza inasible, cosa sumamente estorbosa para proyectos computacionales no lingüísticos que necesitan dar al lenguaje por hecho.

Además, en lo que concierne a la lingüística, es necesario deslindarse de ciertos enfoques que han dejado huella en los estudios del lenguaje de las últimas décadas. Por ejemplo, no es objetivo de análisis de este trabajo —ni siquiera secundario— el hacer juicios acerca de la *gramaticalidad* de alguna estructura sintáctica o morfosintáctica (es decir, si está o no morfológica o sintácticamente bien formada). Eso es un reto que todavía depende de muchas cosas sin resolverse. De hecho, en un corpus representativo de una lengua como el que aquí se ha escogido como fuente primordial de información, no son extrañas las construcciones

[12]No se trata aquí de minimizar la importancia de los aspectos no lingüísticos de la tesis. Por supuesto que es importante asegurarse que un algoritmo no sea tan complejo que nunca llegue a un fin; pero el que, por ejemplo, un programa sea tan complejo o poco "elegante" que tenga que invertir días para contar o examinar de otra manera tal o cual fenómeno lingüístico no será motivo para no llevarlo a cabo.

tanto sintácticamente "mal" formadas como de gramaticalidad cuestionable[13]. Así, como en todo proyecto basado en córpora, el mejor punto de partida no parece ser el de hacer juicios sobre la aceptabilidad gramatical de la evidencia lingüística. De hecho, parece más pertinente medir lo que hay en el corpus, sin juzgarlo. De esta manera, en este trabajo se recurre al cálculo de índices para cuantificar algunas de las propiedades formales de las estructuras allí presentes sin detenerse a calificar dichas estructuras.

Esto está íntimamente ligado a que en este trabajo se renuncia de antemano a las pretensiones teóricas que se quieran hacer acerca de algún esquema o formalismo seleccionado para describir la lengua estudiada. Es decir, aquí se reconoce que utilizar tal o cual esquema descriptivo, no implica de ninguna manera que el lenguaje descrito tenga la misma estructura de dicho esquema (por ej., el utilizar cadenas de Markov no significa que éstas constituyan ni el mejor ni el único esquema descriptivo de las lenguas naturales). Además, en ausencia de información propiamente cognoscitiva, no es propósito de este trabajo describir fenómenos de carácter cognoscitivo.

Por último, otra restricción pertinente es la de dedicar este proyecto a los niveles morfológico y morfosintáctico en su dimensión estructural y no de significación (aunque de ninguna manera se acepta que la gramática sea independiente del significado). De hecho, no todos los temas de la morfología pueden abarcarse en un espacio como éste, así que habrá cuestiones morfológicas (como la parasíntesis) que podrían haberse tocado pero que no se tratarán en este trabajo.

---

[13]Sobra decir que los mejores criterios para hacer estos juicios automáticamente son de carácter estadístico y no dependen tanto de la aplicación de los formalismos lingüísticos más conocidos (compárense Gazdar y Mellish, *op. cit.* [52] 1989, y Charniak, *op. cit.* [32] 1993).

Lo que hay que resaltar es que, tanto el importantísimo nivel fonológico (tan caracterizado en los estudios de córpora no computarizados[14]), como el sintáctico, además de los temas típicos de la semántica y las relaciones estructurales dictadas por el contenido léxico de las palabras o por algún esquema gramatical conocido (orden de palabras, argumentos del verbo, papeles temáticos, etc.) quedan fuera del alcance de este proyecto.

0.4 Metodología

En todo lo anterior está implícita la idea de apoyarse en el corpus como fuente primordial de los hechos lingüísticos, prescindiendo de agregarle manualmente cualquier información adicional producto de la reflexión u observación humana (no aplicable automáticamente), por obvia que pudiera parecer[15]. Esto constituye, de hecho, la decisión de carácter metodológico más importante del proyecto (porque, entre otras cosas, resuelve por sí misma otras cuestiones metodológicas). Esto implica la construcción de diversos programas computacionales[16] para

---

[14]En contraste con esos córpora y aunque el *CEMC* comprende porciones de lengua hablada (transcrita), éste último no proporciona datos propiamente fonológicos. Si acaso se pueden aproximar las letras a signos de carácter fonológico mediante reglas como las de la tabla A.7 de la página 358, pero obviamente esa no es manera de estudiar los fenómenos del nivel fonológico. Sería extraño hacer fonología sin datos fonológicos.

[15]Así, por ejemplo, no se tomaron en cuenta las marcas gramaticales aplicadas al corpus (descritas en el apéndice, página 359), ya que, si bien muchas fueron producto de un procedimiento automático (véase García Hidalgo, "La formalización del analizador gramatical del DEM" [50], en Lara, Ham y García, *op. cit.* [89] 1979), un gran porcentaje fue aplicado manualmente con base en los criterios no siempre consistentes de varios analistas. De todas maneras, las marcas de una porción del *CEMC* ya han servido para construir por lo menos un etiquetador gramatical probabilístico (véase Héctor Jiménez y Guillermo Morales, "SEPE: A POS Tagger for Spanish" [72] en Gelbukh, *Computational Linguistics and Intelligent Text Processing* [53], Springer, 2002, pp. 250-259).

[16]El proyecto se llevó a cabo en diversas computadoras personales mediante un *compilador* del lenguaje de programación C++ de Borland International, *C++ Development Suite for Windows 95, NT, 3.1 and DOS* [18, 21, 20, 19], versión 5.01, 1996. Una excelente introducción a este lenguaje es de Craig Arnush, *Teach Yourself Borland C++ 5* [11], Sams Publishing, Indianapolis, 1996. Una breve discusión de las ventajas y desventajas de los lenguajes de programación más usados o más disponibles para el estudio de los fenómenos lingüísticos está en el primer apéndice "Main Programming Languages and their Suitability" de Barnbrook, *Language and Computers* [13], Edinburgh University Press, Edinburgh, 1998.

explorar y procesar el corpus, así como para extraer de allí y caracterizar cuantitativamente las unidades lingüísticas pertinentes.

No menos importantes son los procedimientos metodológicos característicos de un trabajo de investigación. Primero, se desarrollaron una serie de experimentos para falsificar y refinar cada una de las proposiciones manifestadas en las hipótesis presentadas más abajo. Estos experimentos permitieron *deducir* a partir del corpus un conjunto de afijos (las principales unidades del nivel morfológico estudiadas en este trabajo) y un conjunto mínimo de vocablos gramaticales. Después, mediante un procedimiento de *inducción* lógica que involucra nuestra presunción de representatividad del corpus, se asumió la cualidad de los resultados observados como descriptivos del español de México. Sólo de esta manera se puede argüir que los resultados de esta investigación —deficientes o no— obedecen a por lo menos algunos de los patrones lingüísticos de nuestra lengua.

Otro aspecto metodológico importante es el de los medios y herramientas conceptuales que se utilizan en un trabajo de investigación. Así, este proyecto presupone, aparte de la conveniencia del equipo utilizado[17], herramientas conceptuales cualitativas tales como la lógica o la teoría de conjuntos y cuantitativas tales como las teorías de probabilidad, de información (entropía, cadenas de Markov, etc.) y la estadística de digramas. La validez general de estos medios justifica su uso común en los más diversos trabajos de investigación, dentro y fuera de la lingüística. Esto motiva su aplicabilidad en este proyecto.

[17]Es decir, el objeto material que obviamente consiste en, encarna o aloja las estructuras investigadas, pero también en medios conceptuales tales como lenguaje artificial, compilador, el concepto mismo de computadora, etc.

0.5 Premisas

Pero los procedimientos metodológicos y los medios conceptuales no son los únicos presupuestos de un trabajo de investigación. Éste también descansa sobre la validez de premisas que merecen hacerse explícitas para evitar que pasen desapercibidas presuposiciones no deseadas (premisas tácitas no cuestionadas[18]). Conviene enumerar las ideas, proposiciones o hipótesis, cuya validez se presume en este trabajo. Aunque algunas ya se han mencionado arriba, las más importantes se pueden ordenar de la siguiente manera:

1. Metodología: Los métodos y medios conceptuales mencionados en el apartado anterior (el uso de un corpus, como producto material de una lengua, para la deducción de información que por inducción sea pertinente al objeto infinito que es esa lengua: asimismo, las teorías de conjuntos, de probabilidad, de la información, de estadística de digramas, etc.).
2. Conjeturas o premisas prematuras: este trabajo no se adhiere a las creencias (1) de que la introspección es un método suficiente para identificar o dilucidar datos lingüísticos, y (2) de que un programa computacional es capaz de simular estados o eventos mentales[19]. Más que explicar los resultados de esta investigación, no adherirse a estas conjeturas justifica algunas decisiones metodológicas tomadas.
3. Premisas sobre los sistemas lingüísticos: el concepto de afijo se presta a la investigación cuantitativa; la lengua es un sistema que tiende a la economía de signos; el contenido semántico de una forma influye en la estructura formal del contexto de esa forma, pero no está en esa estructura misma, sino en el proceso de significación.
4. Asunciones sobre el objeto de estudio (que todavía merecen investigarse a fondo[20]): el CEMC es representativo del español de México: los caracteres que conforman a ese corpus corresponden a los fonemas de esa lengua; la palabra española corresponde a la secuencia de caracteres que aparece entre espacios o signos de puntuación; los prefijos españoles no son sintácticamente pertinentes; las categorías gramaticales tradicionales del español (art., adv., adj., etc.) y sus relaciones estructurales en la oración española (por ej., el adj. va junto al sustantivo).

---

[18]Mario Bunge, *Scientific Research I. The Search for System* [24]. Berlín/Heidelberg, Springer-Verlag. 1967. pp. 226-227.

[19]Véase John Searle, "Minds, Brains, and Programs" [124], *The Behavioral and Brain Sciences*, 3(1980). pp. 422-424.

[20]Fórmulas tentativas para permitir una deducción que se retienen o rechazan según la fuerza de sus consecuencias, Bunge, *op. cit.* [24] 1967, p. 226.

5. Problemas de medición: los errores son inevitables (tanto aquellos de aplicación de criterios, como de transcripción de datos, así como aquellos de juicios sobre lo aceptable del dato lingüístico —es decir, de *gramaticalidad*), pero los delatan sus bajas frecuencias (en comparación con las de los fenómenos sistemáticos del lenguaje) y convergen hacia ciertos valores[21] (es decir, se distribuyen normalmente), cosa que no se puede asumir de otros fenómenos, especialmente los lingüísticos, para los que se han observado otras distribuciones.

De entre estas premisas, la referente a la palabra española merece singular atención, especialmente porque la palabra hablada no corresponde exactamente a la palabra escrita. Esto se hace particularmente obvio en el capítulo sobre el clítico. De esta manera, si bien aceptar provisionalmente esta premisa permite, por un lado, empezar el trabajo, por el otro, nos habilita para, paradójicamente, mostrar que las palabras gráficas no pueden considerarse todas del mismo nivel.

Dada esta premisa, a veces me referiré a la palabra en general como palabra gráfica (aunque el corpus contenga porciones transcritas de lengua hablada) sin que se entienda que me refiero sólo a un tipo u a otro. De todos modos, debido a la naturaleza del corpus, todo lo logrado en este trabajo será pertinente sobre todo a la lengua escrita, cosa que como se verá no evita que se abra una ventana hacia el fenómeno hablado.

Con respecto a las otras premisas enumeradas, éstas no agotan todas las presuposiciones de la tesis. Pero son las más importantes porque sirven de punto de partida para la construcción de las hipótesis que se presentan en la próxima sección. Las premisas están ordenadas de lo general a lo específico: medios conceptuales, hipótesis de carácter lingüístico y algunos conceptos característicos del español. Si en efecto estas premisas son ciertas, podemos inferir,

[21] Error *verdadero* o residuo en la terminología de Woods *et al.*, *Statistics in Language Studies* [136]. Cambridge University Press, Cambridge, 1986, pp. 77-93; y error típico en Gonzalvo, *Diccionario de metodología estadística* [57]. Morata, Madrid, 1978, p. 72.

entre otras cosas, que los datos extraídos del corpus en esta investigación constituyen una descripción del español de México.

0.6 Hipótesis

Como se apuntó arriba, en un corpus se encuentran rastros de una multitud de fenómenos lingüísticos. El problema de extraer de allí información de este tipo, además de empírico, es conceptual, metodológico y evaluativo. Así, el concebir unidades, por ejemplo morfológicas, implica determinar una serie de características que las definan. Cuando esos rasgos son de tipo más formal o estructural (y menos de significación o contenido), por lo que son susceptibles de contarse y combinarse cuantitativamente con otros rasgos, debe ser posible verificar (o falsificar) la presunción de que esas características en efecto determinen las unidades lingüísticas examinadas (porque proporcionan criterios para evaluar estas unidades como tales). De esta manera, proponer una solución al aspecto conceptual proporciona una solución al problema evaluativo, mediante la cual se puede resolver el empírico, cosa que a su vez verifica o falsifica la solución al problema conceptual. Las hipótesis sobre *afijalidad* y *cliticidad*[22] que se

[22]Con el término *afijalidad* me refiero a la cualidad que un segmento de palabra (o vocablo) tenga de ser un afijo de la lengua examinada. Este término me parece especialmente afortunado porque el sufijo *~idad*, además de significar 'cualidad de', significa también 'cantidad de' en palabras tales como 'pluviosidad', 'natalidad', 'mortalidad', etc. (véase Rainer, *Spanische Wortbildungslehre* [116], Niemeyer, Tübingen, 1993, p. 530). Así, y como se verá más adelante, el sentido de este término es, aunque cualitativo, en esencia cuantitativo. En lo que respecta a las alternativas para nombrar a este concepto, *afijidad* y *afijabilidad*, el primero va contra la tendencia general del sufijo *~idad* de adherirse a adjetivos (a pesar de las poquísimas excepciones de sustantivos que también son adjetivos, por ej., 'hermandad' y 'complicidad'). El segundo implica un proceso —del verbo 'afijar'— o la posibilidad de su resultado. Esto es mucho más que la simple cualidad de ser afijo. De hecho, el término podría ser útil para referirse a las diferentes medidas de probabilidad que se pueden calcular de un afijo (véase la sección 1.7.2 del primer capítulo). Similarmente, el término *cliticidad* designará a una cantidad indicadora de la cualidad que una palabra tenga de adherirse a otras palabras en sus contextos (es decir, de acompañarlas subordinadamente u orbitar alrededor de ciertos tipos de éstas). La construcción de esta forma puede considerarse paralela a la de 'analiticidad' a partir de 'analítico' ('cliticidad' de 'clítico').

presentan abajo hacen operativos los conceptos de afijo y clítico como elementos formales observables en el corpus. El carácter cuantitativo de estas hipótesis operativas nos permite así evaluar segmentos de palabras y palabras completas en cuanto a sus respectivas naturalezas como afijos o clíticos y aceptarlas o no como instancias de dichos conceptos, lo que proporciona una solución a los tres aspectos del problema de extraer datos lingüísticos del corpus. Además, sus diferencias y semejanzas permiten proponer y examinar la hipótesis de *glutinosidad*[23] —expuesta al final de esta sección— que las generaliza.

Como se estableció arriba, podemos presumir que en el corpus se exhibe una oscilación entre lo esperado y lo inesperado en el discurso y, en relación con esto, entre signos gramaticales y de contenido. Tradicionalmente se ha utilizado la mera frecuencia como método para distinguir entre lo esperado y lo inesperado, así como entre lo que es gramatical y lo que no. También conceptos como la entropía[24] han servido para determinar cuestiones similares, especialmente en cuanto a la estructura de las palabras se refiere (como se dijo arriba, al medir la impredecibilidad de ocurrencia de los segmentos en ciertos contextos se descubren fronteras morfológicas). Provisionalmente, definiremos este concepto sencillamente como el cantidad de información de un conjunto de signos. De esta manera, si medimos la cantidad de información de los segmentos gramaticales, esperaríamos que fuera menor que la de los de

---

[23]Del adjetivo *glutinoso* (lat. *glūten*, 'cola, engrudo'; véase Joan Corominas y José A. Pascual, *Diccionario crítico etimológico castellano e hispánico* [44], Gredos, Madrid, 1991 [1980], *s.v.*, GLUTEN), este término se referirá a una propiedad no tanto *de* los elementos (como la afijalidad y la cliticidad), sino más bien *entre* ellos.

[24]Claude Shannon y Warren Weaver, *The Mathematical Theory of Communication* [125], University of Illinois, Urbana, 1964 [1949], pp. 14, 50. Este concepto también se conoce como 'caos', 'desorganización', 'incertidumbre', etc. y paradójicamente, también como 'información', 'organización', 'sorpresa', etc., porque la medida de los primeros fenómenos (que hasta cierto punto son lo mismo) corresponde a la de los segundos: la información necesaria para organizar el caos es supuestamente del mismo tamaño que la incertidumbre que causa ese caos o que la sorpresa que nos ocasiona. Véase la descripción formal de esta herramienta en el primer capítulo (a partir de la página 83) y su aplicación al *CEMC* en el dedicado al afijo (página 106).

contenido[25]. En otras palabras, cabe esperar que los signos que le dan estructura al léxico y al sintagma —tales como los afijos y los clíticos— contengan menos información que los otros tipos de signos.

El otro aspecto que se mencionó antes con respecto a la estructura lingüística implícita en un corpus se refiere a la capacidad combinatoria de los signos, es decir, a las estructuras de signos que sirven como vehículo del intercambio informativo del discurso. Esto es medible de muchas formas. De hecho, para contar la presencia de estructuras combinatorias en un corpus, éstas se pueden definir de diversas maneras; por ejemplo, contando las combinaciones de signos llamados *cuadrados* o *cuadros*[26]. Por otra parte, está la caracterización de esas combinaciones según su capacidad de generar nuevos signos de los niveles siguientes; por ejemplo, el mayor o menor número de objetos a los que los afijos y clíticos se adhieren es una aproximación a su cualidad económica. Lo importante es que esto, como la entropía y los cuadros, también es susceptible de contarse (mientras se adhieran a más signos, más económicas serán sus relaciones)[27].

---

[25]Por ejemplo, considérese el sufijo *~mente*, cuyo significado —de naturaleza predominantemente gramatical— se captura en un párrafo breve: "Sufijo que sirve para formar adverbios de modo, añadiéndolo a los adjetivos femeninos: 'prontamente', 'sabiamente'. Puede añadirse acomodaticiamente a todos los adjetivos que lo admiten por su significado" (Moliner, *Diccionario de uso del español* [104], *s.v.*). Como se ve, se trata sobre todo de información del contexto en que aparece. Por otra parte, el sustantivo 'mente' —de naturaleza obviamente léxica— es mucho más informativo (con respecto al mensaje donde pueda aparecer) y se define en el mismo diccionario en términos de vocablos de contenido con tantos significados como 'inteligencia', 'facultad', 'pensamiento', 'intimidad', etc. En el *Diccionario del español usual en México* ([90], *s.v.*) no hay todavía una entrada para el sufijo, pero la definición del sustantivo requiere también de términos con alto contenido de información: 'pensamiento', 'inteligencia', 'conciencia', 'juicio', etc. Como veremos adelante, la definición técnica de entropía o sorpresa nos permite estimar esta diferencia de contenido informativo: lo que más nos sorprende contiene más información que lo que nos sorprende menos.

[26]Joseph Greenberg, *Essays in Linguistics* [58], The University of Chicago Press, Chicago, 1957, p. 20. La definición formal de estas estructuras y su aplicación al *CEMC* se examinan con detalle en el capítulo sobre el afijo (página 104).

[27]Un método interesante para medir estas relaciones económicas es el cociente propuesto por Josse de Kock y Walter Bossaert, *op. cit.* [82] 1978, pp. 21-26, 30-32. Las ideas detrás de este método se presentan en el capítulo siguiente (página 87) y su aplicación al *CEMC* en el capítulo sobre el afijo (a partir de la página

De esta manera, si concebimos a los afijos de todas las lenguas (y, de hecho, también a los clíticos) como objetos que se han desgastado fonológica y semánticamente a tal grado que —después de ser objetos plenos e independientes— ahora sólo aparecen adheridos a otros objetos, podemos hipotetizar que —además de ser muy frecuentes y ocurrir, por lo tanto, en un gran número de estructuras combinatorias— contienen poca información en el sentido técnico del término (porque son muy probables en la cadena hablada) y se adhieren a muchísimos objetos para darles estructura a éstos y al discurso en que aparecen. La ventaja de estas propiedades es que son susceptibles de medirse de diversas maneras.

Pero lo importante no es tanto qué métodos se utilizan para medir estas propiedades, sino la relación que éstas guardan las unas con las otras en cuanto a su pertinencia como parámetros para determinar afijos y clíticos. Parece obvio que en ambos casos tanto la frecuencia, como las capacidades combinatorias de los segmentos son directamente proporcionales, mientras que su contenido de información debe ser inversamente proporcional. Sin embargo, al tratarse de propiedades de tipos de objetos (afijos y clíticos) presentes no solamente en el español, sino que también —en mayor o menor medida— en las lenguas del mundo, estaremos en posición de argüir que estas magnitudes abstractas (prescindiendo de los métodos particulares que se puedan aplicar para medirlas) también son pertinentes en la estimación de las propiedades cualitativas que los segmentos de todas las lenguas tengan de ser afijos o clíticos.

---

111).

Hipótesis de *afijalidad*

Los afijos —en su calidad de signos que le dan estructura gramatical a las palabras[28]— contienen menos información (o, al ocurrir en la cadena hablada, sorprenden menos) que las bases con que se asocian, ya que las segundas llevan el grueso del contenido transmitido por el texto. Además, en comparación con estas últimas, los afijos no sólo son considerablemente más frecuentes, sino que también el número de combinaciones en las que participan es mucho mayor. Finalmente, el hecho de ser muy pocos, con respecto al número de palabras que conforman, implica una relación de economía entre los signos del nivel morfológico y los signos del nivel léxico.

En otras palabras, los afijos de una lengua se caracterizan cuantitativamente por su número limitado, su alta frecuencia, sus muchos contextos, su baja entropía (menor contenido de información) y su alta participación en varios vocablos con numerosos segmentos de baja frecuencia (cosa que se traduce en una mayor economía de signos). Esto quiere decir que, para cada segmento posible de cada vocablo de un corpus, se puede calcular un índice *AF* —mediante los promedios de, por lo menos, sus frecuencias, cuentas de contextos, entropías y medidas de economía— que representa la cualidad de ese segmento de ser un afijo de la lengua representada en el corpus. En términos más concretos, se puede postular que la afijalidad de un segmento es directamente proporcional al producto de la cuenta de sus contextos con alguna medida de economía e inversamente proporcional a su entropía: $AF(s_x) = \frac{f_x c_x k_x}{h_x}$, donde f_x, c_x, k_x y h_x representan respectivamente la frecuencia, la cuenta

[28]Los afijos, tanto derivativos como de flexión, codifican información pertinente tanto a la estructura interna de la palabra (son marcas de tipos de palabras —adj., adv., etc.) como a la externa, pertinente al discurso (por ej. fenómenos anafóricos y de concordancia).

de contextos, la economía y la entropía del segmento s_x calculadas a partir de los vocablos de un corpus Ψ.

Hipótesis de cliticidad

De manera similar, los clíticos —en su calidad de signos que le dan estructura al sintagma— contienen menos información (o sorprenden menos al ocurrir en la cadena hablada) que las palabras plenas alrededor de las cuales gravitan. Además, en comparación con estas últimas, los clíticos son considerablemente más frecuentes y, por lo tanto, el número de combinaciones en las que aparecen es mucho mayor. Por último, el hecho de ser muy pocos, con respecto al número total de vocablos, implica una relación económica entre los signos del nivel del sintagma y las frases posibles del nivel siguiente.

Así, los clíticos de una lengua se caracterizan cuantitativamente por su número limitado, su alta frecuencia, sus muchos contextos, su baja entropía (menor contenido de información) y su alta ocurrencia junto a numerosos vocablos de baja frecuencia. Esto significa que para cada vocablo de un corpus se puede calcular, según sus contextos, un índice *CL* —mediante, cuando menos, su frecuencia, su número de contextos y sus medidas de entropía y economía— que cuantifica la cualidad que pueda tener el segmento de ser un clítico con respecto a los otros vocablos con los que ocurre en el corpus. Concretamente, la cliticidad de un vocablo es directamente proporcional al producto de su frecuencia y alguna medida de economía e inversamente proporcional a su entropía: $CL(v_x) = \frac{f_x c_x k_x}{h_x}$, donde f_x, c_x, k_x y h_x representan respectivamente la frecuencia, el número de contextos, la economía y la entropía del vocablo v_x calculadas a partir de un corpus Ψ.

Hipótesis de glutinosidad

Tanto la hipótesis de afijalidad como la de cliticidad están planteadas en términos de ciertas dimensiones calculables en cada segmentación al interior de cada vocablo o entre los vocablos o presuntos vocablos que ocurren adyacentes en un corpus dado. A saber —y como ya se dijo arriba—, para que un segmento sea afijal, se espera una segmentación económica, con un alto número de cuadros y una baja entropía. De manera similar, la cliticidad es también una función de estas dimensiones, pero entre segmentos gráficamente independientes (es decir, aquellos entre espacios). Entonces, podemos concebir una medida de pegajosidad o glutinosidad entre estos segmentos, idónea en la caracterización de las hipótesis de afijalidad y cliticidad.

Así, al interior de los vocablos, esta pegajosidad alcanzaría su valor máximo entre bases y afijos y, a su exterior, alcanzaría su valor mínimo (aproximándose a cero) entre estructuras discursivas complejas (períodos, párrafos, textos). En otras palabras, al interior de la raíz de un vocablo, cabría esperar una alta entropía e índices mínimos de economía y cuadros; y entre bases y afijos, una menor entropía y mayores índices de economía y cuadros; mientras que entre clíticos y palabras, la entropía podría ser todavía menor (pero mayor que entre sintagmas) y las medidas de economía y cuadros serían todavía mayores (aunque menores que entre sintagmas). Por lo que la glutinosidad entre un afijo a_x y una base b, sería directamente proporcional a la afijalidad del primero: $GL_{a_x::b} \approx AF(a_x) = \frac{f_x \bar{c}_x \bar{k}_x}{\bar{h}_x}$, donde f_x, $\bar{c}_x$, $\bar{k}_x$ y $\bar{h}_x$ representan respectivamente la frecuencia, el número de cuadros (o combinaciones), la economía y la entropía del segmento a_x calculados a partir de un corpus Ψ. Por otra parte, al exterior de la palabra la pegajosidad debe ser también directamente proporcional a la cliticidad (a

mayor cliticidad, mayor glutinosidad, porque menos cliticidad implica más independencia entre segmentos o, incluso, una posible frontera entre sintagmas): $GL_{c_x::v} \approx CL(c_x) = \frac{f_x \bar{c}_x \bar{k}_x}{\bar{h}_x}$. donde f_x, $\bar{c}_x$, $\bar{k}_x$, $\bar{h}_x$ son la frecuencia, los cuadros, la economía y la entropía del segmento clítico c_x con respecto al corpus Ψ.

A continuación y para terminar esta sección, se resumen las hipótesis que sirven de espina dorsal al desarrollo de trabajo. La primera hipótesis corresponde al capítulo segundo, la segunda se examina en el tercero y la última es tema del capítulo cuarto:

1. La afijalidad de un segmento de palabra se puede caracterizar cuantitativamente mediante la fórmula: $AF(s_x) = \frac{f_x c_x k_x}{h_x}$, donde f_x, c_x, k_x y h_x representan respectivamente la frecuencia, la cuenta de contextos, la economía y la entropía del segmento s_x calculadas a partir de los vocablos de un corpus Ψ.

2. Similarmente, la cliticidad de una palabra se puede caracterizar cuantitativamente mediante la fórmula: $CL(v_x) = \frac{f_x c_x k_x}{h_x}$, donde f_x, c_x, k_x y h_x representan respectivamente la frecuencia, los contextos, la economía y la entropía del vocablo v_x calculadas a partir de un corpus Ψ.

3. De lo anterior se desprende que la afijalidad y la cliticidad son dos casos de propiedades de unidades lingüísticas que corresponden a una misma fuerza, posibilidad de enlace o glutinosidad entre los segmentos de un corpus Ψ, que se puede caracterizar cuantitativamente mediante la fórmula: $GL(s_x) = \frac{f_x \bar{c}_x \bar{k}_x}{\bar{h}_x}$, donde f_x, $\bar{c}_x$, $\bar{k}_x$ y $\bar{h}_x$ representan respectivamente la frecuencia, las combinaciones, la economía y la entropía del segmento s_x en relación a dicho corpus.

Así, en el capítulo sobre el afijo determinaremos los valores de estos parámetros para cada segmento de cada vocablo del corpus con el objeto de constatar que son factores pertinentes al carácter de afijo que dichos segmentos puedan tener (por ej. si el segmento 'ando' es más sufijo en 'tom~ando' que en 'cu~ando'). En el capítulo sobre el clítico se calcularán los mismos valores para las palabras gráficas con el objeto de constatar si también son factores pertinentes en el descubrimiento de clíticos (por ej. si el pronombre 'me' es más clítico que

el verbo 'caracterizar' o que el adverbio 'cerca'). Naturalmente, una manera de aproximarse a estas hipótesis es encontrar evidencia a favor, pero también en contra de ellas, es decir, mostrar que los segmentos gramaticales (afijos y clíticos) ni llevan menos información, ni son menos sorprendentes, ni más económicos, ni ocurren en más estructuras combinatorias que los segmentos de contenido. Eso en sí sería un logro importante.

Por último, se examinará la relación entre ambos tipos de segmentos para construir una generalización de ambos fenómenos (¿es lo que hay entre la base 'trabaj~' y el sufijo '~ado' de la misma naturaleza que aquello que hay entre el clítico 'se' y el verbo finito 'durmió'?).

0.7 Plan de la tesis

En este último apartado se describe el contenido de la tesis que consta de esta introducción, cuatro capítulos, las conclusiones y los apéndices. En esta introducción se hicieron los planteamientos generales de este trabajo de investigación: se delimitaron los objetivos, se justificaron las decisiones metodológicas para llevarlos a cabo, se presentaron las hipótesis, se hicieron explícitas las premisas y se precisaron algunos conceptos de importancia al desarrollo de la tesis.

El primer capítulo presenta el panorama general en el que se enmarca el trabajo de esta investigación. Primero se describen algunos enfoques de carácter automático que se ocupan de describir los fenómenos inherentes al exterior de la palabra gráfica, es decir sintácticos. Luego se examinan algunos métodos automáticos de descubrimiento de signos léxicos (y en especial de sus *colocaciones*) y de signos gramaticales dentro del sintagma. Finalmente se

examinan los esquemas más conocidos de la morfología automática y, especialmente, las técnicas de segmentación morfológica de palabras.

En el capítulo segundo —"El afijo en el *CEMC*"—, se presenta la investigación del nivel morfológico, destinada a llevar a cabo los primeros dos objetivos de la tesis, es decir, determinar automáticamente un conjunto de signos al interior de la palabra y construir un programa segmentador de vocablos a partir de ese conjunto. Para eso, se presenta una discusión de procedimientos de descubrimiento de unidades morfológicas, y se determina al afijo como el tipo de unidad morfológica más apta de ser investigada automáticamente. Después se hace una investigación empírica de los afijos en el *CEMC*, para falsificar la hipótesis sobre afijalidad de segmentos de palabras que se describió arriba. Por último, se presentan los resultados.

El tercer capítulo —"El clítico en el *CEMC*"— se ocupa de los objetivos tercero y cuarto de la tesis, es decir, de la investigación cuantitativa al exterior de la palabra gráfica. En esencia se examinan los métodos explorados en el capítulo anterior para descubrir afijos. También se examina el fenómeno de la puntuación como indicio de fronteras sintagmáticas. Finalmente se aplican dos de esos métodos y un índice de puntuación para determinar los vocablos gráficos que más se adhieren a otros vocablos. De esta manera, la investigación empírica en este capítulo se centra en el descubrimiento de clíticos a partir del *CEMC*: se busca falsificar la hipótesis sobre la cliticidad de cadenas de caracteres que ocurren entre dos espacios. Al final se presentan los resultados.

En el capítulo cuarto —"Glutinometría en el *CEMC*"—, se presenta un esquema generalizador de las herramientas para descubrir afijos y clíticos que se investigaron en los capítulos anteriores. Se trata de establecer en qué consiste la relación entre afijalidad y cliticidad

y expresar ambas cualidades en términos de una fuerza de asociación entre segmentos del discurso, que se puede concebir como una especie de pegamento estructural o glutinosidad. Pero una fuerza de asociación de este tipo merece construirse en el marco de la teoría de la medición. De esta manera, el problema se analiza en términos de las posibles dimensiones y unidades de la glutinosidad. Además, se determinan los axiomas para una medición de esta fuerza, es decir, para una glutinometría. Por último se presentan los resultados de la aplicación de este sistema de medición al *CEMC*.

En las conclusiones se hace una recapitulación de los logros de la tesis. se examinan las ventajas y desventajas de los métodos aplicados y se reformulan las hipótesis presentadas en esta introducción.

En el primer apéndice —"El *Corpus del Español Mexicano Contemporáneo*"— se describen brevemente algunas de las características generales del *CEMC*. Se incluyen apartados sobre su preprocesamiento, sus formas (o vocablos) más frecuentes y las abreviaturas encontradas automáticamente. El segundo —"Muestra aletoria de vocablos analizados"— es esencialmente la lista de los vocablos del *CEMC* escogidos al azar para comparar los índices de segmentación automática de palabras examinadas en el capítulo sobre el afijo. En forma tabular aparecen marcados para cada vocablo los aciertos de sólo los valores más altos de cada índice. El tercer apéndice —"Sufijos en el *CEMC*"— es el catálogo de los 749 fragmentos de palabra más sufijales del corpus según los criterios aplicados también en el capítulo sobre afijos. Por último, en el apéndice —"Las formas más gramaticales del *CEMC*"— se consignan los 500 vocablos gráficos *gramaticalmente* más pertinentes, según los criterios de economía y entropía caracterizados a lo largo de la tesis.

Capítulo 1

Panorama general

El presente capítulo contiene los antecedentes generales de la investigación llevada a cabo en este trabajo. Primero se exploran brevemente las bases del análisis sintáctico automático, esto es, del exterior de la palabra gráfica. El poco espacio no permite examinar con profundidad cada uno de los enfoques típicos de esta amplia área de estudio. por lo que apenas se mencionan algunos formalismos, escuelas y trabajos específicos. Se reserva más espacio para explorar algunos métodos de descubrimiento de unidades léxicas (*lexical acquisition*) y procedimientos para medir la asociación entre palabras gráficas. típicamente aplicados en la determinación de unidades fraseológicas y *colocaciones*. La segunda parte del capítulo se ocupa del interior de la palabra, es decir, del nivel morfológico. Se trata de una breve revisión bibliográfica de algunos trabajos dentro de la morfología automática. especialmente de aquellos que se ocupan de la segmentación de palabras en morfemas.

1.1 Análisis automático al exterior de la palabra

En esta sección se toca brevemente el panorama de los estudios que utilizan computadoras para investigar fenómenos lingüísticos al exterior de la palabra gráfica. Esto es importante

para construirnos un idea general de las investigaciones automáticas de este tipo. De esta manera, se hará un recuento de los formalismos gramaticales más conocidos, empezando por los fundamentos que planteó Noam Chomsky.

A pesar de las limitaciones que el formalismo transformacional ha encontrado en la implementación de sistemas de generación y análisis computarizados, dentro de la lingüística automática, es en los estudios del nivel sintáctico donde quizá más mella ha hecho la escuela generativista de Chomsky[1]. Como es bien sabido, la contribución chomskiana a la teoría de autómatas asistió al desarrollo de las matemáticas discretas y la computación. Sin embargo, ya que esta doctrina sintáctica, en sus diferentes formulaciones, no trascendió tal cual como el enfoque más apto de aplicarse en la construcción de sistemas informáticos, pronto aparecieron otras doctrinas sintácticas casi siempre también de corte universalista y mentalista (HPSG, LFG, etc., que se comentarán brevemente adelante).

Una de las reflexiones más importantes de Chomsky se refiere al contraste entre lo infinito del lenguaje y el carácter necesariamente finito de los córpora, cosa que hace evidente la necesidad de estudiar el lenguaje con otros recursos. Esto motivó, sin embargo, cierta impopularidad de la estadística en la investigación lingüística[2], cosa que se refleja en las doctrinas sintácticas herederas del transformacionalismo. De hecho, a partir de los años cincuenta se le cerró al *mainstream* lingüístico las puertas de una importante herramienta del quehacer científico en otros campos del conocimiento y que en la lingüística sobrevivió solamente dentro

[1] Algunos trabajos fundamentales del pensamiento de Chomsky son *Estructuras sintácticas* [39], Siglo XXI, México, 11ª ed., 1994; "Finite State Languages" [36], *Information and Control*, 1(1958), pp. 91-112; *Cartesian Linguistics* [35], University Press of America, 1966; Chomsky y Miller, "Introduction to the Formal Analysis of Natural Languages" [37], y "Finitary Models of Language Users" [38], en Luce, Bush y Galanter, *Handbook of Mathematical Psychology II* [92], 1963, pp. 269-322 y 419-492.

[2] Manning y Schütze, *op. cit.* [93] 1999, pp. 4-5.

de ciertas áreas del estudio del habla, como la sociolingüística y los estudios principalmente lexicográficos basados en córpora. Sin embargo, más y más estrategias de tipo estadístico se han hecho más y más populares en ciertos estudios y aplicaciones de fenómenos lingüísticos, incluso en aquellos dedicados al estudio del sistema, especialmente en el campo de la sintaxis automática[3].

Análisis automático y tipos de gramáticas formales

En esta subsección se hace una breve presentación de las ideas detrás del análisis formal de la sintaxis, según se lleva a cabo en la lingüística computacional. Entre los conceptos básicos, se examinan la jerarquía chomskiana de lenguajes de estructura de constituyentes restringidos y los modelos conocidos como redes de transición.

El gran interés por investigar automáticamente los fenómenos lingüísticos al exterior de la palabra se inició fuera de la lingüística, en el marco de la inteligencia artificial, bajo el rubro de procesamiento del lenguaje natural. Aunque el atractivo principal no era la investigación propiamente lingüística, sino la investigación de la inteligencia en general, era natural que buscaran en la lingüística las herramientas conceptuales necesarias para llevar a cabo sus estudios[4]. Y si bien es cierto que la inteligencia artificial no se ha caracterizado por su empirismo (finalmente no estudian la inteligencia natural[5]), también es cierto que la doctrina

[3]Por ejemplo, los trabajos de desambiguación sintáctica se han beneficiado mucho de la aplicación de información estadística en sistemas de análisis automático. Descripciones de varias técnicas probabilísticas aplicadas al análisis del lenguaje se encuentran en Allen, *op. cit.* [4] 1995; y Charniak, *op. cit.* [32] 1993.

[4]No fue así desde el principio. Por ejemplo, hubo numerosos proyectos de traducción automática en que se consideraba que mediante la mera traducción de vocablos y el simple reordenamiento de éstos se lograrían traducciones adecuadas. Los desafortunados resultados de estos y otros trabajos tempranos (por ej. los de interrogación de sistemas —*question answering systems*) hicieron obvio que era necesario tomar en cuenta a los estudios del lenguaje como un fenómeno pertinente a sus objetivos.

[5]Véase, por ejemplo, Gevarter, para quien —en su introducción a la inteligencia artificial— 'empírico'

dominante en la lingüística de las últimas décadas fue la sintaxis generativa, con un enfoque mentalista más bien distante de esquemas verdaderamente empíricos (en cuanto a "estados de hecho" se refiere[6]). Finalmente, si la inteligencia artificial no estudia al lenguaje, pero requiere de un artefacto que le permita simularlo, el enfoque chomskiano sirvió su cometido.

Así, aunque indudablemente hay muchas maneras de investigar al lenguaje, los ingenieros de la inteligencia artificial adoptaron sin mucho problema el método de formular estructuras de información gramatical, especialmente para analizar automáticamente la sintaxis, sobre todo del inglés. Era natural que lo último que les interesara fuera investigar a la gramática misma: la preocupación era cómo representarla, no estudiarla, mucho menos concebirla como un problema[7].

Una de las diferencias principales entre la programación de computadoras en el marco de la computación en general y aquella en el de la inteligencia artificial es que el 'conocimiento' (las reglas gramaticales en los trabajos asociados al lenguaje) que se utiliza en las aplicaciones queda implícito en las instrucciones que configuran al programa en la primera, mientras que separado y codificado en estructuras simbólicas en la segunda[8]. La primera se refiere a la representación procesal (*procedural*) y la segunda se refiere a la representación declarada

---

parece ser sinónimo de 'tomado de otro lado', es decir, sacado del cuerpo de conocimientos de otra disciplina, no de la realidad observable (en el marco de 'sistemas expertos' se puede codificar el conocimiento mediante 'asociaciones empíricas' que a menudo ocultan relaciones causales, por lo que generalmente se prefiere representar el conocimiento de manera 'más profunda', mediante otras asociaciones estructuradas —por el analista— para reflejar la estructura, función y causalidad del conocimiento, véase *Inteligente Maschinen* [54], VCH, Weinheim, 1985, p. 67).

[6]Recuérdese definición en Abbagnano *op. cit.* [1] 1991. *s.v.*, EMPÍRICO.

[7]Se puede argüir que comprobar que una gramática razonada rinde cuenta (o no) de la realidad lingüística es una actitud empírica, es decir, de observar los hechos reales. Sin embargo, al seguir esta dinámica es natural que el interés se desplace del complicado fenómeno lingüístico al formalismo con el que se trabaja: de la ciencia fáctica a la ciencia formal.

[8]Gevarter *op. cit.* [54] 1985, p. 4.

(*declarative*)[9]. De esta manera, los sistemas de lenguaje natural cuentan por lo general con una estructura simbólica declarada[10] donde reside la información gramatical, que alguien saca "empíricamente" de algún lado, en el mejor de los casos de los libros de gramática[11] o de un hipotético hablante ideal.

A un programa o autómata que simplemente determine si una expresión es o no gramatical, se le conoce como reconocedor o aceptador. Uno que, además de determinar si las expresiones son o no gramaticales, proponga una estructura descriptiva de éstas es un analizador gramatical o *parser*, esto es, un programa que infiere estructuras a partir de reglas gramaticales y cadenas de caracteres o, en otras palabras, un aparato que toma una gramática y una cadena de palabras y determina la estructura gramatical de esa cadena (aquella que le permite al analizador hacer juicios sobre la gramaticalidad de dicha cadena[12]).

Así, un formalismo o gramática es un sistema formal que define la membresía de la colección de expresiones lingüísticas y es la herramienta que le sirve al analizador para asignar a cada miembro una estructura y una interpretación. La gramática es simplemente una definición abstracta del conjunto de objetos lingüísticos bien formados. Es un lenguaje construido para describir a otros lenguajes —al conjunto de oraciones que los abarcan (cadenas

---

[9]Gazdar y Mellish, *op. cit.* [52] 1989, pp. 4-5.

[10]Antes de los años ochenta también había aplicaciones donde la información gramatical quedaba implícita en las instrucciones, por ejemplo, la versión original del programa SPANAM (primero de varios programas de traducción entre las lenguas oficiales de la Organización Panamericana de la Salud —inglés, francés, portugués y español; véase Slocum, "Machine Translation: A Survey of Active Systems" [128] en István S. Bátori, Winfred Lenders and Wolfgang Putschke, eds., *Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications* [14], Walter de Gruyter, Berlín/Nueva York, 1989, pp. 637-638).

[11]Así, Hallebeek (*A Formal Approach to Spanish Syntax* [61], tr. Pieter de Haan, Editions Radopi, Amsterdam/Atlanta, 1992) consulta las gramáticas tradicionales del español: Bello, Gilli Gaya, etc.

[12]Fernando C. N. Pereira, y Stuart M. Shieber, *Prolog and Natural-Language Analysis* [112], Center for the Study of Language and Information, Stanford, 1987, p. 35.

de caracteres), sus propiedades estructurales (sintaxis) y, presuntamente, sus significados (semántica). El formalismo gramatical, "metalenguaje" según Shieber, codifica el análisis del lenguaje objeto. Es una herramienta para la descripción de lenguajes naturales: delimita la clase de lenguajes naturales posibles; y proporciona una caracterización de lenguajes naturales "interpretable" por una computadora[13].

En este marco se pueden concebir varios tipos de lenguajes y de gramáticas que los describan. Chomsky, al explorar la relación entre la teoría de autómatas y los fenómenos gramaticales, propuso una tipología jerárquica para los lenguajes de estructura de constituyentes restringidos (*constituent-structure grammars*)[14], que dan su nombre al tipo de gramática formal o sintaxis que los describe mejor[15]:

Tipo 0. Lenguajes recursivamente enumerables,

Tipo 1. Lenguajes sensibles al contexto,

Tipo 2. Lenguajes independientes del contexto (*Context-Free*),

Tipo 3. Lenguajes regulares.

Los lenguajes de tipo cero tienen una sintaxis sin límite y se describen mediante sistemas de reescritura no restringida, por ejemplo una máquina de Turing[16]. Para Chomsky estos

---

[13]Stuart Shieber, *An Introduction to Unification-Based Approaches to Grammar* [126], Center for the Study of Language and Information, Stanford, 1986 (CSLI Lecture Notes, 4). p. 5.

[14]Noam Chomsky, "On Certain Formal Properties of Grammars" [33], *Information and Control*, 2 (1959). pp. 137-167; y "Formal Properties of Grammars" [34] en R. D. Luce, R. Bush y E. Galanterm, eds.. *op. cit.* [92] 1963. Esta jerarquía aparece por lo general en los manuales de lingüística computacional; por ejemplo. en Hopcroft y Ullman, *Introduction to Automata Theory, Languages, and Computation* [69], Addison-Wesley. Reading (Mass.), 1979, pp. 217-232; en Naumann y Langer, *op. cit.* [108] 1994. p. 17; en Gazdar y Mellish. *op. cit.* [52] 1989, pp. 132-141 o en Peter Hellwig, "Parsing natürlicher Sprachen: Grundlagen" en I. S. Bátori. W. Lenders y W. Putschke, eds., *op. cit.* [14] 1989, p. 356.

[15]No se trata aquí de gramáticas del tipo transformacional por el que aboga.

[16]Una máquina de Turing pueden considerarse "nothing more or less than a program of a perfectly arbitrary kind for a digital computer with potentially infinite memory", Chomsky art. cit. [34] 1963, pp. 359-360.

sistemas son de poco interés lingüístico porque, si bien mediante una máquina de Turing se puede definir un procedimiento finito y bien definido como lo sería la descripción automática de un lenguaje, se trata de un sistema demasiado abierto (no lo suficientemente restringido) para parecerse a una gramática de lenguaje natural, a no ser que ilumine los rasgos estructurales que distingue a los lenguajes naturales de los conjuntos de símbolos arbitrarios y recursivamente enumerables. De allí los tipos de lenguajes siguientes que Chomsky define imponiendo sucesivamente condiciones restrictivas.

De esta manera, si las gramáticas para lenguajes de tipo cero contienen reglas de reescritura del tipo $\phi \rightarrow \psi$ donde ϕ y υ son cualquier cosa o conjunto de cosas, aquellas para los del tipo uno requieren que el número de símbolos de la izquierda sea menor o igual al de la derecha y que ningún símbolo sea nulo: $a_1...a_m \rightarrow b_1, ..., b_n$ donde $m \leq n$. Estos lenguajes también son recursivos, pero los contextos limitan la aplicación de las reglas de las gramáticas que los describen[17].

Para los lenguajes de tipo dos se agrega la restricción de requerir un sólo símbolo no terminal en el lado izquierdo de todas las reglas de las gramáticas que los describen: $a \rightarrow b_1, ..., b_n$, es decir, los símbolos están en cierta manera "aislados" de sus contextos con respecto a los símbolos con que se reescriben[18].

Por último los lenguajes de tipo tres se caracterizan por tener una sintaxis de estados finitos. Se trata de lenguajes que son simplemente demasiado restringidos para rendir cuenta de lenguas naturales.

---

[17] *Ibid.* [34], p. 360. Nótese que a más símbolos a la izquierda, mayor contexto involucrado.

[18] *Ibid.* [34], pp. 366-410, donde se explora este tipo de gramáticas.

Lo importante de esta tipología y de las reflexiones que de allí se desprenden es que todo esto ha servido de fundamento para una multitud de aplicaciones y trabajos de investigación lingüística basados en la formulación de reglas, sobre todo a nivel sintáctico[19]. Desde el punto de vista generativista hay muchas razones para no utilizar autómatas de estados finitos en la construcción de gramáticas para lenguajes naturales. De hecho, tampoco cualquier máquina abstracta que pueda describir lenguajes recursivamente enumerables (como una máquina de Turing) sería adecuada. Al tratarse de sistemas demasiado abiertos, supuestamente se acaba el interés lingüístico. Como veremos a continuación, eso no ha sido motivo para que nuevas escuelas de la sintaxis automática no se basen en esquemas que describen lenguajes del tipo 0 (como las redes de transición aumentadas) para construir herramientas y sistemas *teóricos* de interés lingüístico.

Otros formalismos gramaticales

En este subapartado se examina la gama de formalismos que han sido creados como herramientas descriptivas del lenguaje natural y que aspiran a analizar e incluso generar cadenas de texto bien formadas. En la búsqueda de lenguajes formales capaces de analizar lenguas naturales, los lingüistas se dividen en dos grupos con diferentes prioridades: aquellos que diseñan herramientas lingüísticas y aquellos que construyen *teorías* lingüísticas. Por un lado, los criterios de diseño que utilizan los primeros son comparativamente independientes de los análisis lingüísticos subyacentes a las herramientas que construyen. Por el otro, los crite-

[19] A menor escala también en el morfológico. Más adelante, en este capítulo, se examinan los transductores de la fonología de estados finitos, muy conocidos dentro de la morfología computacional, aunque en realidad se ocupan poco de lo morfológico (a partir de la página 52).

rios de los segundos están íntimamente ligados a *sus* análisis lingüísticos, porque sus criterios mismos pretenden encarnar principios lingüísticos universales[20] (mentalistas y universalistas).

Sin embargo, ambos tipos de enfoques formales se hacen operativos mediante un mismo esquema formal, conocido como de redes de transición aumentadas (*augmented transition networks ATN*) que William A. Woods propuso originalmente[21]. Estas redes son básicamente redes de transición recursivas equipadas con memoria y con la habilidad de incrementar (*augment*) el número de arcos entre estados con acciones y condiciones que hacen referencia a esa memoria[22] (lo que las hace capaces de describir lenguajes del tipo 0). Las gramáticas formales que se derivaron de este tipo de red utilizan dicha memoria para albergar "cartas de navegación" (*charts*), estructuras de rasgos. etc. Las "cartas de navegación" sirven para que el analizador ordene sus metas del momento o sus intentos afortunados o fallidos al analizar subconstituyentes de la cadena de palabras. Las estructuras de rasgos, por otra parte, son el lugar donde el analizador coloca la información sobre la cadena de palabras que va determinando durante su ejecución.

Los formalismos que dependen de alguna operación o estrategia para combinar (*unify*) estas estructuras de rasgos son llamados de "unificación" (*unification-based*), pero también han sido designados como de "información o restricción" (*information- or constraint-based*) porque presuponen que la información gramatical de una frase es discreta o restringida. Es decir, las propiedades de las frases simplemente están o no están especificadas y si lo están.

---

[20]Shieber, *op. cit.* [126] 1986, p. 38.

[21]William Woods, "Transition Network Grammars for Natural Language Analysis", *Communications of the ACM*, 13:10 (1970), pp. 591-606.

[22]Gazdar y Mellish, *op. cit.* [52] 1989. p. 6.

las posibilidades son discretas (el número de una frase es singular o plural, un verbo puede requerir un sujeto ya sea animado o inanimado. etc.)[23].

Hay muchos esquemas con diferentes grados de popularidad. Entre ellos se encuentran las gramáticas léxico-funcionales, funcionales de unificación, de cláusulas definidas, generalizadas de estructura de frase. Algunos esquemas que siempre se mencionan en los manuales de este tipo de lingüística computacional son:

- Joan Bresnan y Ronald Kaplan desarrollaron su gramática léxico-funcional (*Lexical-Functional Grammar LFG*) a partir del concepto de red de transición aumentada y del trabajo de Bresnan sobre lingüística no transformacional de orientación léxica. Este tipo de gramática es básicamente una gramática libre de contextos y se postula como teoría lingüística.
- Martin Kay ideó un formalismo gramatical funcional —gramática funcional de unificación (*Functional Unification Grammar FUG*)— para el diseño de herramientas lingüísticas.
- Alain Colmerauer produjo los sistemas-q y la gramática de metamorfosis (*Metamorphosis Grammar*).
- Pereira y Warren la gramática de cláusulas (oraciones) definidas (*Definite Clause Grammar DCG*) que también fue pensada como herramienta lingüística y a su vez ha dado lugar a otros formalismos.
- Gazdar desarrolló la gramática generalizada de estructura de frase (*Generalized Phrase Structure Grammar GPSG*)[24] que se postula como teoría lingüística. U. Klenk la ha aplicado al español.
- Pollard y Sag definieron la gramática de núcleo (*Head Grammar*) que dio lugar a la gramática de frase verbal dirigida por núcleos (*Head-Driven Phrase Structure Grammar HPSG*).

En realidad hay muchas más gramáticas y enfoques que los que aquí podemos mencionar. La idea central es que todos están basados en la dinámica de formular reglas para describir

[23]Stuart M. Shieber. *Constraint-Based Grammar Formalisms. Parsing and Type Inference for Natural and Computer Languages* [127], MIT Press, Cambridge (Mass.)/London, 1992. p. 16.

[24]Gerald Gazdar, Edwan Klein, Goeffrey Pullum e Ivan A. Sag, *Generalized Phrase Structure Grammar* [51], Harvard University Press, Cambridge (Mass.), 1985.

lenguas naturales, como si esa fuera la única forma de investigarlas. A pesar de la diversidad de enfoques y opiniones sobre su validez como investigaciones lingüísticas, estos esquemas se han aplicado con relativo éxito sobre todo en la industria. Sin embargo, es indudable que hay otras formas de investigar el lenguaje. Específicamente, la que incumbe al presente trabajo es, como se dijo en la introducción, una basada en la investigación cuantitativa de un corpus que, sin capturar el carácter ilimitado de un lenguaje natural, sí constituye una muestra estadística de la población infinita que ese lenguaje constituye. Finalmente, nada impide que estas maneras de investigar se complementen.

1.2 Descubrimiento de vocablos gramaticales

Si el corpus es una muestra del habla que a la vez refleja y se produce a partir de lo gramatical, una manera de empezar a investigar los asuntos gramaticales es determinar automáticamente los signos que injieren en la estructura de la lengua, es decir, de descubrir mediante métodos cuantitativos las palabras o vocablos de uso gramatical (el objetivos 3 y 4 de la tesis). En esta sección se revisan los posibles criterios para determinar cuáles palabras gráficas asumen una función gramatical en la lengua de un corpus, es decir, los procedimientos para medir su carácter estructural (porque le dan estructura al discurso), en oposición al carácter *pleno* de las palabras de contenido. Así, a continuación se analizan métodos que van desde la simple determinación de frecuencias, hasta el cálculo de diversos índices estadísticos.

1.2.1 Las frecuencias de los vocablos

En esta subsección se examina la frecuencia como criterio para determinar el carácter gramatical de un vocablo. En particular, revisaremos el procedimiento para determinar, a partir del mismo *CEMC*, los vocablos más frecuentes del español que merecieran conformar la nomenclatura del DEM. Desde el principio fue claro que al tratarse de un léxico abierto, un corpus representativo serviría para determinar solamente las palabras más usuales de la lengua y no la totalidad del léxico usado en México. Pero se juzgó acertadamente que la frecuencia absoluta de los vocablos no era lo más apropiado. Hubo que matizar el criterio de frecuencia para dar cuenta de los diversos géneros representados heterogéneamente (con diferentes tamaños)[25] en el corpus.

Según la notación utilizada, en el *CEMC* hubo Ω vocablos (v_i, donde $1 \leq i \leq \Omega$) y se tomaron en cuenta m géneros (G_j, donde $1 \leq j \leq m$). Se construyeron automáticamente tablas siguiendo el modelo de la tabla 1.1. Allí se almacenaron, primero, las frecuencias absolutas de cada vocablo en el corpus (t_i, donde $1 \leq i \leq \Omega$) y las frecuencias absolutas de cada uno en cada género (f_{ij}, donde $1 \leq i \leq \Omega$ y $1 \leq j \leq m$) y, luego, las frecuencias relativas de cada vocablo con respecto a sus ocurrencias entre los géneros (e_{ij}) y con respecto a las ocurrencias de los otros vocablos dentro cada género (d_{ij}). También se estimó el tamaño relativo de cada género (r_{ij}):

$$e_{ij} = \frac{100 f_{ij}}{t_i}, \quad d_{ij} = \frac{100 f_{ij}}{\sum_{j=1}^{\Omega} f_{ij}}, \quad r_j = \frac{100 \sum_{i=1}^{\Omega} f_{ij}}{\sum_{i=1}^{\Omega} t_i}$$

[25] Sobre la cuestión de la heterogeneidad del *CEMC*, véase Lara, "La cuantificación en el DEM" [87], en *op. cit.* [85] 1990, pp. 56-68. Los géneros en que está dividido el *CEMC* aparecen en la tabla A.1 del apéndice (página 351).

Luego se calculó una frecuencia corregida para cada vocablo (KF[26]), solamente a partir de las frecuencias absolutas y el tamaño relativo de los géneros:

$$\mathrm{KF}_i = \frac{\left(\sum_{j=1}^{m} \sqrt{r_j f_{ij}}\right)^2}{100}$$

y un índice (C_i) normalizado (con valores entre 0 y 1) para indicar la dispersión de cada vocablo entre géneros (1 significa distribución uniforme en todos los géneros):

$$C_i = \frac{100S_i - \min_i r_j}{100 - \min_i r_j}, \quad \text{donde} \quad S_i = \frac{\mathrm{KF}_i}{t_i}$$

De esta manera, con una frecuencia corregida y un índice de dispersión para cada vocablo, se seleccionó una lista de los más usuales en el español de México.

Tabla 1.1: Análisis estadístico de los vocablos del *CEMC*

	frecuencias absolutas					**frecuencias relativas**								**medidas estadísticas**		
						$\frac{100t_i}{\sum_{k=1}^{\Omega} t_k}$	entre géneros				dentro de géneros						
	total	G_1	G_2	...	G_m		G_1	G_2	...	G_m	G_1	G_2	...	G_m	KF	S	C
v_1[a]	t_1	f_{11}	f_{12}	...	f_{1m}	h_1	e_{11}	e_{12}	...	e_{1m}	d_{11}	d_{12}	...	d_{1m}	KF_1	s_1	
v_2	t_2	f_{21}	f_{22}	...	f_{2m}	h_2	e_{21}	e_{22}	...	e_{2m}	d_{21}	d_{22}	...	d_{2m}	KF_2	s_2	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
v_i	t_i	f_{i1}	f_{i2}	...	f_{im}	h_i	e_{i1}	e_{i2}	...	e_{im}	d_{i1}	d_{i2}	...	d_{im}	KF_i	s_i	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
v_Ω	t_Ω	$f_{\Omega 1}$	$f_{\Omega 2}$	...	$f_{\Omega m}$	h_Ω	$e_{\Omega 1}$	$e_{\Omega 2}$	...	$e_{\Omega m}$	$d_{\Omega 1}$	$d_{\Omega 2}$	...	$d_{\Omega m}$	KF_Ω	s_Ω	

[a]Tomado de Lara y Ham, art. cit. [88] 1974 en Lara *op. cit.* [89] 1979, pp. 36.

Naturalmente el criterio de frecuencia es también un criterio probabilístico. Escoger los más frecuentes es escoger los más probables. Y los más probables deben ser los que organizan el discurso y no tanto los que transmiten su contenido. De hecho, las probabilidades reales de éstos últimos deben ser menores que las que se estiman a partir del corpus porque, si bien en un corpus de buen tamaño esperaríamos la ocurrencia de todos los signos gramaticales, sólo ocurriría una porción de los de contenido (al considerar que constituyen una clase abierta).

[26] *Korrigierte Frequenz* creada por J. Lanke.

La tabla A.9 que aparece en la página 362 del apéndice reproduce la lista de los 100 lemas más usuales según el criterio de frecuencia descrito arriba. Aparte de los lemas de los verbos y algunos sustantivos muy frecuentes en todos los géneros, la mayoría corresponde a las palabras que tradicionalmente conocemos en español como gramaticales. El índice de dispersión, por otra parte, no es un indicador confiable para determinar signos gramaticales, principalmente porque algunos de ellos se distribuyen en géneros determinados. Por ejemplo, los pronombres 'yo' y 'me' que tienen dispersiones comparativamente bajas (31 y 41 en la tabla A.9), ya que tienden a ocurrir más en conversaciones y en literatura, que en géneros tales como los científicos o técnicos. De hecho, se puede hipotetizar que el que ciertos signos de este tipo ocurran en ciertos géneros y no en otros debe ser señal de "alguna característica cualitativa de los textos incluidos en ellos"[27].

Obsérvense también las 188 formas sin lematizar con frecuencias absolutas más altas en la tabla A.10 del apéndice (página 365). La tabla 1.2 reproduce las primeras 20. De nuevo, con excepción de ciertos sustantivos y adjetivos muy frecuentes (algunos de los cuales se podrían considerar casi gramaticales), aparece una gran mayoría de los signos que se conocen como gramaticales en la tradición española. Mientras recorremos la lista de más a menos frecuentes, aparecen formas como [kreo] o [pesos] que seguramente tienen una dispersión baja entre los diversos géneros. Pero lo importante aquí es establecer que, independientemente de la finura de los cálculos de frecuencias, el criterio de seleccionar las más frecuentes resultará en la selección de las más gramaticales (entre las que, de no tener una medida de dispersión que las elimine, podrán encontrarse palabras léxicas mucho muy frecuentes en unos pocos

[27]Lara art. cit. [87] 1990, p. 66.

Tabla 1.2: Las formas sin lematizar del *CEMC* con frecuencias absolutas mayores

núm.	forma[a]	fr. absoluta	grafías
001	[de]	118879	de. dé
002	[la]	75963	la (art.. pron.)
003	[ke]	71801	que. qué
004	[i]	65356	y
005	[a]	58983	a. ha
006	[el]	56771	el. él
007	[en]	51951	en
008	[se]	35658	sé. se
009	[los]	31985	los (art.. pron.)
010	[no]	31362	no
011	[las]	21364	las (art.. pron.)
012	[un]	20277	un
013	[por]	20054	por
014	[es]	19681	es
015	[del]	19168	del
016	[kon]	19072	con
017	[una]	16669	una (art.. pron.)
018	[o]	15422	o
019	[para]	14796	para
020	[si]	14716	sí. si

[a]No se tomaron en cuenta los acentos, se eliminaron las abreviaturas y se aplicaron las modificaciones descritas en la tabla A.7.

géneros).

De todas maneras, es importante enfatizar que la gramaticalidad de un vocablo corresponde a más que un bajo contenido de información determinado por una alta frecuencia. por lo que las frecuencias no son de ninguna manera suficientes para determinar qué tan gramatical es una forma. También podrían considerarse otros criterios más elaborados como la productividad de las formas que, como veremos más adelante en este capítulo —en la discusión de productividad morfológica de Baayen a partir de la página 56—. también es susceptible de medirse cuantitativamente. Pero, como veremos. la productividad no se corresponde completamente con una gramaticalidad medida mediante frecuencias: si hemos de concebir las formas gramaticales como las más productivas (cosa también cuestionable). encontraremos, como Baayen, que "productivity and token frequency are independent. Un-

productive classes may encompass far more tokens than productive ones"[28].

1.2.2 Otros métodos de adquisición léxica

Como con el *CEMC*, una de las aplicaciones más conocidas de los córpora es la determinación de las unidades léxicas que merezcan pertenecer a los diccionarios. En el marco de otros córpora, se han propuesto otros métodos estadísticos conocidos bajo el rubro de 'adquisición léxica' (*lexical acquisition*). En esta subsección se examinan algunos de éstos que, concretamente, se han utilizado para medir la asociación entre vocablos y que, por lo tanto, se podrían aplicar al descubrimiento de palabras gramaticales. Este rubro comprende diversos métodos con diferentes objetivos, entre los cuales están:

- determinar tipos de digramas, trigramas, etc., que constituyan unidades fraseológicas, colocaciones (*collocations*), sintagmas y otras combinaciones de palabras (marcos de subcategorización verbal),
- aclarar ambigüedades en los sentidos de las palabras (*word-sense disambiguation*), resolver ambigüedades de dependencia de frases preposicionales (*PP attachment disambiguation*),
- determinar similitudes semánticas entre palabras.

En los manuales de estadística de córpora[29] no hay procedimientos estrictamente de descubrimiento de unidades gramaticales en general, pero, cuando menos, se encuentran siempre métodos de adquisición léxica (que no tienen que ver con los fenómenos de adquisición

[28]Baayen. *A Corpus-Based Approach to Morphological Productivity* [12]. Academisch Proefschrift, Centrum voor Wiskunde in Informatica, Amsterdam, 1989, p. 50.

[29]Véanse, por ejemplo, Manning y Schütze, *op. cit.* [93] 1999, y Oakes, *op. cit.* [111] 1998.

del lenguaje[30], sino con la determinación de propiedades sintácticas y semánticas de las palabras[31]). En general, el objetivo principal de estos métodos es el desarrollar algoritmos y técnicas estadísticas para examinar la ocurrencia de patrones de palabras en córpora de gran tamaño. A continuación analizaremos los métodos de determinación de unidades fraseológicas, que miden la asociación entre palabras (los otros métodos tienen como objetivo el esclarecimiento de estructuras, más que la determinación de asociación).

El problema de definir la unidad fraseológica

Aunque el término 'colocación' puede definirse de varias maneras[32], en los manuales de estadística de córpora se utiliza para designar aquellas unidades fraseológicas que son expresiones formadas por dos o más palabras y que corresponden a una manera convencional de decir algo, es decir, se define como dos o más palabras consecutivas con un comportamiento peculiar, que se comportan como una unidad sintáctica o semántica[33].

El principal criterio[34] para determinar si un grupo de palabras constituye una colocación

[30]Manning y Schütze, *op. cit.* [93] 1999, apuntan que, aunque con estos métodos de adquisición léxica no se intenta construir un esquema de la adquisición del lenguaje en el ser humano, el éxito que logran tiende a socavar los argumentos clásicos chomskianos que postulan un lenguaje innato basados en una presunta pobreza de los estímulos, p. 265.

[31]Los métodos de adquisición léxica son métodos para construir y complementar diccionarios cuantitativo-electrónicos y otras herramientas.

[32]Según Firth, "Collocations of a given word are statements of the habitual or customary places of that word" ("A Synopsis of Linguistic Theory 1930-1955", p. 181, citado por Manning y Schütze, *op. cit.* [93] 1999, p. 151). También pueden verse como grupos de palabras (sustantivos, verbos, adjetivos) que tienden a coincidir en estructuras sintagmáticas aun cuando sus partes no aparecen de manera consecutiva (por ej. sal y pimienta, clavo y martillo, gato y ronronear) e, incluso, como la distribución de vocablos que tienen algún morfema en común (por ej. leer, lectura, lector).

[33]Véase Manning y Schütze, *op. cit.* [93] 1999, pp. 183-186.

[34]Hay otros criterios, tales como la traducción de una lengua a otra: si no se puede traducir palabra por palabra, se presume que se trata de una colocación. Manning y Schütze dan como ejemplo la traducción literal

es la ausencia de composicionalidad, porque el significado de una unidad fraseológica de este tipo no es la composición de los significados de sus partes. Además, las palabras que la constituyen no se pueden modificar libremente ni cambiar por otras.

Obviamente esto es un problema en el marco de los procedimientos automáticos, porque para determinar la ausencia de composicionalidad se requiere, por ahora, de la participación activa de un analista capaz de tomar decisiones cualitativas sobre la naturaleza de asociación entre palabras. Hay toda una serie de procedimientos que se han aplicado en la determinación de las llamadas colocaciones[35]. Algunos muy conocidos son, además del criterio de frecuencia y cálculos de promedios y varianzas, la prueba de χ^2 (de Pearson), razón de semejanza (*likelihood ratio*) e información mutua. De hecho, el éxito relativo de estos métodos nos debe hacer suponer que una asociación estadísticamente importante entre dos elementos corresponde a un cierto desgaste de su significado composicional: a más asociación estadística, más desgaste composicional.

Frecuencia

El método más sencillo para determinar colocaciones es simplemente el ordenar los digramas presentes en un corpus de los más a los menos frecuentes. Cuando dos palabras ocurren juntas con frecuencia, hay lugar para sospechar que al aparecer juntas tengan alguna función distinta a la que se esperaría que podría resultar de su combinación.

---

de la frase inglesa *make a decision* al francés **faire une décision* (que se traduce propiamente como *prendre une décision*) como evidencia de que la primera es una colocación. Pero no queda claro si la colocación no es la frase francesa, sobre todo al tomar en cuenta que en otras lenguas es diferente, por ejemplo, en alemán las decisiones no se toman, sino se encuentran: *eine Entscheidung treffen*.

[35] Manning y Schütze, *op. cit.* [93] 1999, pp. 153-183.

Manning y Shütze seleccionaron los digramas más frecuentes (frecuencia absoluta) de un corpus de cuatro meses de las noticias por cable del periódico *New York Times*[36]. La tabla 1.3 reproduce los resultados. Aunque para ellos esta lista de digramas no es muy interesante.

Tabla 1.3: Digramas más frecuentes en un corpus del *New York Times*

frecuencia[a]	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11429	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been

[a]Sacado de Manning y Schütze, *op. cit.* [93] 1999, p. 154.

podemos observar —como era de esperarse— que los digramas más frecuentes (con excepción de '*New York*') son pares de palabras función o gramaticales. Estos resultados no les parecen interesantes, porque están tratando de hacer exactamente lo contrario de lo que se intenta hacer en el presente trabajo: quieren descubrir construcciones léxicas o de contenido y no gramaticales. De hecho, recurren a un filtro o *stop-list* (que otros investigadores normalmente aplican en tareas similares) para eliminar las palabras función. Por lo general, este tipo de filtro se aplica también en los métodos que a continuación se describen.

[36]Manning y Shütze, *op. cit.* [93] 1999, p. 153.

Promedios y varianza

Hay pares de palabras que, aunque están asociadas cuantitativamente, no siempre ocurren una después de la otra, sino que otras palabras aparecen entre ellas. Una estrategia es la de definir una ventana de entre tres o cuatro palabras a cada lado de la palabra-objeto examinada. Luego se pueden examinar aquellas palabras que en promedio aparezcan con más frecuencia dentro de la ventana alrededor del objeto. También es posible registrar las distancias entre las palabras y calcular su promedio y varianza.

Así, el promedio y la varianza (o, aun, la desviación estándar) caracterizan la distribución de las distancias entre las palabras examinadas. Según Manning y Schütze, aquellos pares con una desviación estándar baja, lo que implica una distancia más o menos fija (un valor de cero delata una distancia fija), son candidatos a ser colocaciones[37]. Una varianza o desviación estándar alta indicaría que hay una tendencia de una palabra a aparecer aleatoriamente en todas las posiciones de la ventana definida alrededor de la otra, por lo que infieren que la relación entre una y otra no es tan interesante.

Medidas estadísticas

Otros métodos estadísticos se han aplicado para medir la asociación entre palabras. A continuación, se examina la aplicación, en la determinación de colocaciones, de la prueba de t y se introduce la prueba de independencia de χ^2 que se presentará con detalle más adelante en este capítulo, junto con las estadísticas de razón de semejanza e información mutua.

Debido a su aplicabilidad en muestras pequeñas, la prueba de t ha sido ampliamente usada

[37]Manning y Schütze, *op. cit.* [93] 1999, p. 159.

para determinar la asociación entre palabras. Esta prueba requiere del cálculo de promedios y varianzas de los datos recogidos. Como en la prueba de χ^2 (véase la página 73), en la de t se comparan los valores observados con valores esperados, pero aquí se comparan los promedios y su diferencia se divide entre el error estándar[38]:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

donde μ es el promedio de valores de la población, $\bar{x}$ y s son, respectivamente, el promedio y la desviación estándar de los valores de la muestra y N es el tamaño de dicha muestra. Si esta estadística es lo suficientemente grande (cosa que se determina al compararla con un valor extraído de una tabla, según cierto margen de error), se presume que las palabras comparadas están asociadas.

El problema es que para aplicar esta prueba, es imperativo suponer que los datos están distribuidos normalmente y, como lo han señalado diversos estudiosos del lenguaje[39], es muy raro que haya fenómenos lingüísticos que se puedan describir mediante una curva normal.

Por eso, la prueba de independencia de χ^2 es mucho más apropiada para medir la asociación entre palabras. El problema con esa prueba es que no es apropiada para muestras pequeñas —como es el caso cuando se buscan colocaciones— o cuando alguna de las celdas

---

[38]Véanse Woods, Fletcher y Hughes *op. cit.* [136] 1986, pp. 101-103; Manning y Schütze, *op. cit.* [93] 1999, pp. 163-166; y Gonzalvo *op. cit.* [57] 1978, pp. 141-142.

[39]Por ejemplo, esa es la crítica que siempre se hace a libros como el de Oakes (*op. cit.* [111] 1998) que asume que a mayor número de datos, más proximidad a la distribución normal de los fenómenos lingüísticos. Véanse Church y Mercer [40], "Introduction to the Special Issue on Computational Linguistics Using Large Corpora", *Computational Linguistics*, 19 (1993), p. 20; y la reseña [7] que Altmann hace del libro de Oakes en *Journal of Quantitative Linguistics* 6:3(1999), p. 270: "Though it can be shown mathematically that all distributions (under certain conditions) and deviations asymptotically approach the normal distribution, in language it is not so. Here, the greater the sample size the more 'skewed' its properties can become, which is in agreement with the principles of synergetic linguistics, the principle of self- organization (evolution), that of diversification, etc. but not with the principles of asymptotic tests".

de la tabla de contingencia que se construye para calcular la estadística alcanza un número menor a cinco.

La descripción de esta prueba se presenta más adelante junto con las estadísticas para la razón de semejanza y la información mutua (véanse la tabla de fórmulas 1.8 y la de contingencia 1.7), en la parte de este capítulo dedicada a la morfología, ya que en este trabajo se aplicaron en ese nivel lingüístico.

1.3 Morfología automática

Antes de examinar los métodos para medir la asociación entre objetos del nivel morfológico, conviene hacer un breve resumen de lo que se puede llamar morfología computacional, es decir, del estudio mediante computadoras de los fenómenos lingüísticos al interior de la palabra. En este apartado se presentan algunos de los métodos más conocidos de este campo. Esto es importante para formarse un panorama global de los trabajos de investigación automática de los fenómenos morfológicos en general. Más concretamente, la morfología computacional es aquella rama de la lingüística computacional dedicada al estudio del análisis y síntesis (generación o producción) de palabras[40]. El análisis de las palabras implica el reconocimiento de formas implícitas (o *subyacentes*) a partir de las palabras flexionadas o derivadas; es decir, la identificación de los segmentos radicales con respecto a los afijales o a otros radicales. La síntesis, por otra parte, implica la construcción o producción automática de formas derivadas a partir de bases radicales y, especialmente, de formas flexionadas a partir de los lexemas.

[40]Koskeniemmi, "Computational Morphology" [84] en Bright [22], pp. 291-293.

Pero el objetivo principal de la morfología computacional es adquirir un mejor y más explícito entendimiento de la morfología en general. Está de más decir que esto está muy relacionado con el análisis y la producción de oraciones por un lado y con el reconocimiento y la síntesis automáticos del habla (*speech recognition and synthesis*), por el otro.

Su articulación con el análisis sintáctico está íntimamente ligada a la aplicación de etiquetas gramaticales a las palabras según ocurran en textos. Por eso muchos sistemas de análisis gramatical tienen como componentes centrales las descripciones morfológicas de las lenguas particulares que analizan.

Como ya se notó desde la introducción, el inglés es una lengua de flexión muy sencilla, por lo que es costumbre en los sistemas de procesamiento de esa lengua (la mayoría) ignorar este aspecto[41] y simplemente incluir en el componente léxico todas las formas flexionadas y derivadas (estrategia muy poco económica). Por eso, los trabajos más interesantes de morfología computacional se ocupan por lo general de lenguas como las romances, el ruso, el finlandés, etc.

Se ha propuesto que el análisis y generación de palabras dentro del marco de la morfología computacional incluya las siguientes tareas[42]:

1. Estudiar los procesos fonológicos y morfofonológicos que causan variación en fonemas (o letras en la lengua escrita). Esto corresponde al dominio de la fonología.

2. Identificar morfemas de bases y afijos (prefijos, infijos, sufijos, etc.).

---

[41] De hecho, aunque la derivación en inglés es más compleja que su flexión, también tiende a ignorarse, tal vez porque un algoritmo tan sencillo como el de Porter (que veremos a partir de la página 70) basta para desnudar las palabras inglesas de sus afijos flexivos y derivativos, cosa que no se puede decir de lenguas con derivación más compleja.

[42] Koskenniemi, art. cit. [84] 1992, p. 291.

3. Describir las estructuras morfotácticas posibles, es decir, las secuencias y combinaciones posibles de morfemas (la manera en que se combinan prefijos, raíces y sufijos para formar palabras completas).

4. Identificar los rasgos morfosintácticos y las descripciones semánticas de las palabras completas a partir de las descripciones de los morfemas que la forman.

5. Describir el proceso de lexicalización, donde ciertas configuraciones de morfemas tienen propiedades que no pueden deducirse de sus componentes individuales.

El punto número dos —la identificación o descubrimiento de morfemas— es quizá el más descuidado en la morfología automática, a pesar de que en tiempos preinformáticos haya sido objeto de mayor atención[43]. Este punto es el más importante en esta investigación, ya que uno de sus objetivos es determinar un conjunto de morfemas del español de México. Por esta razón, la sección siguiente de este capítulo (sobre segmentación automática de palabras) está dedicada a los métodos más importantes de descubrimiento de morfemas. Mientras tanto, se reseñan dos tipos de estudio automático basados en la formulación y estudio de reglas descriptivas de los fenómenos morfológicos (incluso fonológicos).

El asunto medular de estos enfoques es cómo representar el conocimiento y su estructura. Por ejemplo, los primeros que a continuación se examinan dependen de la noción de regla para esto (fonología de estados finitos). De hecho, el conocimiento del analista de la morfología se representa mediante la formulación de reglas: cuando un sistema de reglas *funciona* hay cierta presunción de que el conocimiento codificado en ellas es pertinente al fenómeno. Anderson apunta:

---

[43]Aun mucho después de la introducción de las computadoras al estudio del lenguaje, las técnicas de descubrimiento de unidades lingüísticas siguen sin recibir la atención que se merecen. Entre las posibles explicaciones para este estado de cosas, está —además de que como, se dijo, sea el inglés (con su sencilla y conocidísima morfología) la lengua de estudio privilegiada— la escasez de córpora electrónicos para la mayoría de las lenguas, especialmente las no europeas.

> the virtue of a computer is that it knows nothing other than what it has been told, and so when one reaches a point at which a computational procedure operates correctly, it is reasonably certain that all of the underlying assumptions and subprocedures involved in the description have indeed been made fully explicit[44].

Para Anderson, esto es aceptable en sistemas cuyo objetivo no es investigar la morfología, pero no es una justificación de su estudio científico mediante computadoras. El problema es que rara vez se revisan las cuestiones que se asumen y aunque éstas no sean muy centrales al trabajo, los teóricos no se pueden dar el lujo de dejar de investigarlas. La condición que Anderson impone para aceptar esta dinámica como científicamente válida es que la naturaleza de las reglas que se formulen refleje los principios de la *teoría* lingüística. Y considera que la morfología cuenta con una gran ventaja en esto, ya que confía en que la vasta literatura rebosante de *conocimientos* de la psicología y de la *ciencia* cognoscitiva[45] se ha constituido en parte de tal *teoría* (cosa que se habrá de probar si no se quiere caer en un optimismo excesivo).

En los trabajos que se comentan a continuación la formulación de reglas es la metodología privilegiada para el 'desarrollo de formalismos lingüísticos'. La presunción más importante es que mediante estos formalismos se pueden construir "precise and elegant descriptions of morphological phenomena"[46].

---

[44] Anderson, *A-Morphous Morphology* [8], Cambridge University Press, Cambridge, 1994, p. 375.

[45] *Ibid.* [8].

[46] Ritchie, *et al.*, *Computational Morphology* [120], The MIT Press, Cambridge, Mass., 1992, p. 11.

1.3.1 Fonología de estados finitos

En este apartado se examina la morfología computacional derivada de los trabajos de Martin Kay y Kimmo Koskenniemi. Aunque de ninguna manera son hoy en día los esquemas más conocidos de la fonología propiamente dicha, se trata de una familia de métodos a menudo conocida bajo el término de 'fonología de estados finitos', porque está basada en la aplicación de un tipo de máquina abstracta llamada en inglés *transducer* (transductor). Estos son los esquemas más populares dentro de la lingüística computacional porque la posibilidad de aplicar estas máquinas facilita la construcción de sistemas computacionales.

A principio de los años ochenta, en el marco de la fonología de reglas generativas, se concibieron los primeros transductores de estados finitos para el reconocimiento y generación de la estructura superficial de las palabras (en oposición a la forma léxica o *subyacente*). Un transductor es una máquina de estados finitos que —como los aparatos que se usan en física para transformar algún tipo de señal (movimiento, onda, excitación, etc.) en otro tipo de señal— sirve para transformar cadenas de unos caracteres en cadenas de otros[47]. Un transductor consiste en, además de un número finito de estados, un conjunto finito de símbolos de entrada, otro de símbolos de salida y la especificación de correspondencias (*mapping*) entre los primeros y los segundos[48]. En cierta manera, los transductores definen dos lenguajes que son traducciones el uno del otro. En la fonología de estados finitos los lenguajes corresponden a los niveles léxico y de superficie de una misma lengua. Así, dada una forma del primer nivel

[47] *Ibid.* [120], p. 19. En otras palabras, se trata de autómatas cuyas transiciones tienen asociadas operaciones de "salida", véase Glück, *Metzler Lexikon Sprache* [56], Verlag J.B. Metzler, Stuttgart, 2000, *s.v.* AUTOMAT.

[48] Véase definición formal en Aho y Ullmann, *The Theory of Parsing, Translation, and Compiling* [2], Prentice-Hall, Nueva York, 1972, p. 224.

se llega a la forma superficial, y dada una forma superficial se puede llegar al nivel léxico o *subyacente*. En la figura 1.1 se ilustra un transductor de estados finitos para una regla de este esquema[49].

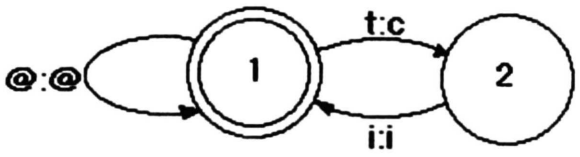

Figura 1.1: Transductor de estados finitos para la regla t:c ⇒ _i:i

Los círculos son estados posibles. El número 1 es simultáneamente el estado inicial y final. Los arcos muestran las equivalencias de símbolos entre niveles (el del nivel léxico antes de los dos puntos ":" y el del superficial después) que se especifican en la regla "t:c ⇒ _i:i", que significa: "If ever the correspondence t:c occurs, it must be followed by i:i"[50]. Es decir, /t/ del nivel léxico podrá manifestarse como [c] en el nivel superficial sólo si ocurre antes de /i/ del nivel léxico que se manifiesta también como [i] en el superficial.[51]. Esta notación contempla también el uso de contextos más complejos (al permitir que se especifiquen tipos de símbolos, por ej., vocales o fricativas alveolares) y de fronteras morfológicas, que se manifiestan en las formas léxicas mediante el símbolo "+". Por ejemplo, considérense las siguientes representaciones que Antworth utiliza para ilustrar la aplicación de las reglas[52]:

---

[49]Tomado de Antworth, *PC-KIMMO: A Two-level Processor for Morphological Analysis* [10]. Summer Institute of Linguistics, Dallas, 1990, p. 44.

[50]*Ibid.* [10].

[51]El arco "@:@" se refiere a todos los demás símbolos que describen estos dos niveles de la lengua en cuestión y se especifica para permitir el paso de otros caracteres por esta regla.

[52]Antworth, *op. cit.* [10] 1990, p.31.

LR (*lexical representation*):	0	t	a	t	+	i
	↕	↕	↕	↕	↕	↕
SR (*superficial representation*):	'	t	a	c	0	i

donde "0" es el símbolo nulo que cancela símbolos de un nivel que no ocurran en el otro; "'" es el acento silábico del nivel superficial; y "+" es, como se dijo arriba, la frontera morfológica (que corresponde al símbolo nulo en el nivel superficial). El procesador analiza cada par de caracteres de un nivel y otro para determinar si se corresponden según el conjunto de reglas que conforma la descripción de la lengua. En este ejemplo, el programa avanza de izquierda a derecha hasta llegar a la cuarta columna donde al aplicar la regla de la tabla 1.1. verifica que 't' del nivel léxico se manifiesta como 'c' en el superficial cuando aparece antes de 'i' de los dos niveles (sin importar la frontera entre morfemas). Además, el procesador se puede hacer funcionar como *reconocedor* o como *generador*. En el primero acepta representaciones superficiales para obtener las formas léxicas (por ej.[53], *bigger* → 'big+er. *spies* → 'spy+s. *foxes* → 'fox+s). En el segundo acepta las representaciones *subyacentes* y genera las formas superficiales ('fox+s+'s → *foxes'*, 'cat+s → *cats*, 'try+ed → *tried*).

La investigación lingüística más citada sobre la relación entre reglas fonológicas y transductores de estados finitos se debe a Martin Kay y Ronald Kaplan, data de principio de los ochenta y nunca se publicó[54]. Kay y Kaplan sugirieron que las reglas ordenadas de la fonología generativa se pueden hacer computacionalmente operativas mediante una secuencia 'en cascada' de transductores de estados finitos (*cascading sequence of finite state transducers*)[55] y

[53]Nótese que se trata de ejemplos que consideran que la forma superficial es la escrita.

[54]Para discusiones de este trabajo, véanse Ritchie *et al.*, *op. cit.* [120] 1992, p. 20; Koskennieni, art. cit. [84] 1992, p. 292; y Sproat, [131]. *Morphology and Computation*, The MIT Press, Cambridge (Mass.), 1992.

[55]Antworth, *op. cit.* [10] 1990, p. 6.

que dicha cascada siempre puede combinarse para formar un solo autómata. Esto inspiró el modelo de dos niveles de Koskenniemi que se distingue del modelo generativo en que, como se puede apreciar en el ejemplo de arriba, no consta de secuencias de reglas fonológicas que conduzcan de un nivel a otro, sino de reglas que en un solo paso transforman cadenas de un nivel directamente en cadenas del otro (de allí el nombre de *two-level rules*, reglas de dos niveles).

Así, el esquema de estados finitos se distingue del enfoque generativista por las restricciones que se imponen a las reglas para evitar que alguna opere en los resultados de otra[56]. Como se vio arriba, todas las reglas involucran a los dos niveles y son declaraciones lógicas que definen las correspondencias aceptables entre las dos representaciones (por ej., *taci* ↔ 'tat+i se corresponden —una se reconoce en la otra y la otra genera a la primera— de una sola vez mediante "t:c ⇒ _i:i").

Antworth apunta que, además de ser un esquema más económico que el de la fonología generativa (una sola regla de éste corresponde a varias de esta última), es también uno compatible con las ideas de la fonología natural, que rechaza el poder arbitrario e ilimitado de las reglas ordenadas y los resultantes niveles intermedios. El método de Koskenniemi fue divulgado por Karttunen[57] con el nombre de KIMMO, gracias a la publicación de una implementación en LISP acompañada de descripciones de dos niveles del inglés, rumano, francés y japonés[58]

---

[56] Ritchie *et al.*, *op. cit.* [120] 1992, p. 21.

[57] "KIMMO: a General Morphological Processor", *Texas Linguistic Forum* 22(1983), pp. 163-186.

[58] La implementación de Antworth (*op. cit.* [10] 1990) para la computadora PC fue auspiciada por el Instituto Lingüístico de Verano, data de la segunda mitad de los años ochenta y está acompañada de descripciones de muchas otras lenguas.

Es claro que un enfoque de estados finitos no rinde cuenta de todos los fenómenos morfológicos de todas la lenguas. De todas maneras, hay quienes consideran que es suficiente por lo menos para caracterizar lenguas complejas y aglutinantes como el finlandés (aparte de Koskenniemi y seguidores). Jäppinen, por ejemplo, construyó junto con M. Ylilammi un analizador morfológico de estados finitos (diferente al esquema de dos niveles de KIMMO) para esta lengua que se aplicó comercialmente en los ochenta[59]. Este analizador genera una descripción gramatical (marcas de caso, número, posesión, etc.) de las palabras a partir de reglas formuladas por el investigador utilizando su conocimiento del finlandés.

1.3.2 Medidas estadísticas de productividad de reglas morfológicas

En este subapartado se presenta la aplicación de técnicas estadísticas en córpora para investigar patrones morfológicos. En concreto, se examina el trabajo de R.H. Baayen que investiga cuantitativamente el fenómeno de productividad de morfemas o, más específicamente, de reglas de formación de palabras.[60]

Baayen concibe tres componentes de la productividad morfológica que se pueden medir a partir de un corpus[61]: productividad *strictu sensu*, utilidad pragmática de una regla y potencialidad pragmática de esa regla[62].

---

[59] Jäppinen, "Finite State Computational Morphology" [71], en Klenk, *op. cit.* [75] 1992, pp. 96-109.

[60] En términos estrictos son las reglas, y no los morfemas, las que merecen o no la etiqueta de productivas. Para facilitar la exposición, a veces se hablará de productividad de los morfemas, pero quedará implícito que se trata de la productividad de las reglas que involucran a dichos morfemas.

[61] Baayen, *op. cit.* [12] 1989, pp. 25-26.

[62] Baayen ejemplifica todo esto mediante un análisis detallado de la productividad de varios sufijos del holandés, en especial de dos que son "rivales" y sirven para formar sustantivos a partir de adjetivos (*~te* vs. *~heid*). Todo esto le permite a Baayen proponer una interpretación psicolingüística del fenómeno de productividad.

La primera se refiere a la disponibilidad de una regla para aplicarse en un momento dado y se mide mediante el cociente del total de *hapax legomena* (n_1 = total de formas de una sola ocurrencia) sobre el total de ocurrencias de palabras en el corpus en las que se aplica la regla de formación de palabras ($N = \sum_r rn_r$ es la suma de los productos de los totales de formas de todas las frecuencias por esas frecuencias). Así, esta productividad es una estimación de la probabilidad de seleccionar al azar un tipo nuevo a partir del corpus y por lo tanto una medida de productividad de una regla de formación de palabras[63]:

$$\mathcal{P}^{(N)} = \frac{n_1}{N}$$

Por ejemplo, en la tabla 1.4 aparecen los datos que Baayen calculó para algunos sufijos holandeses y las reglas que los involucran. Allí puede constatarse que los sufijos más productivos, según Baayen, tienen los valores de $\mathcal{P}^{(N)}$ más altos (los menos productivos aparecen al final separados por una línea):

Tabla 1.4: Productividad *strictu sensu* y utilidad pragmática de reglas morfológicas en la formación de sustantivos holandeses

	N	n_1	V	$\frac{n_1}{N}$	Z
simplex NOUN	37836	294	1495	0.008	—
~tje	2580	654	1031	0.253	33.2164
~ing	8049	302	943	0.038	24.2983
~heid	2251	256	466	0.114	19.4438
~schap	265	29	64	0.109	11.7109
~sel	261	21	44	0.080	5.5631
~te	758	10	39	0.013	1.9439
~nis	461	6	28	0.013	1.2019

La utilidad pragmática de una regla se refiere a qué tanto se usa esa regla en el corpus. Este aspecto de la productividad se expresa en términos del número de formaciones derivadas mediante una regla dada con respecto al total de ocurrencias de palabras en el corpus, esto

[63] *Ibid.* [12], pp. 55-60.

es, el número de afijos-tipo (*types of an affix*) en un corpus es una medida de su utilidad pragmática:

$$\mathcal{U}^{(F)} = V^{(N)}$$

donde F es el tamaño del conjunto de tipos creados por la regla en cuestión y $V = \sum_{r=1} n_r$ (la suma de los totales de formas de todas las frecuencias de ocurrencia en el corpus) representa el número de tipos en el corpus. En la tabla 1.4 también se especifican los valores de V para los sufijos holandeses que allí aparecen. Nótese de nuevo que los valores más altos de V corresponden a los más productivos (aunque las cantidades de $\mathcal{P}^{(N)}$ exhiban cierta discrepancia, como en el caso de *~tje* y *~ing*, siendo la del segundo. 0.038, considerablemente menor que la del primero. 0.253).

La idea es poder determinar si una regla es productiva, mediante la comparación de su índice de productividad *strictu sensu* con el del conjunto de palabras simples (*simplex*)[64] de la misma categoría, ya que las segundas por definición no son productivas. Así, mientras mayor sea ese índice para una regla dada (y el sufijo asociado a ésta) con respecto al de las palabras *simplex* de la misma categoría, mayor productividad puede presumirse de dicha regla. Baayen muestra que las diferencias entre las reglas productivas y las no productivas son significativas mediante estadísticas del tipo siguiente:

$$Z = \frac{\frac{n_1}{N} - \frac{n'_1}{N'}}{\sqrt{varianza(\frac{n_1}{N} - \frac{n'_1}{N'})}}$$

donde $\frac{n'_1}{N'}$ representa la productividad *strictu sensu* de las palabras *simplex*. Véase la tabla 1.4 para los valores Z de los sufijos que allí aparecen (el renglón '*simplex* NOUN' contiene los valores de n'_1 y N' utilizados en el cálculo de Z). Se puede ver en esa tabla que los procesos

[64]Morfemas libres o palabras no derivadas con las que se pueden construir nuevas palabras.

productivos obtienen valores positivos significativamente altos comparados con aquellos de los no productivos. En otras palabras, al tomar en cuenta los datos de palabras *simplex*, puede verse que la probabilidad de encontrarse nuevas palabras con un sufijo determinado es significativamente mayor para procesos productivos e igual o menor para las clases no productivas[65].

Por último, la potencialidad pragmática se refiere a la potencialidad de una regla de formación de palabras en la matriz socio-cultural. Detrás de todo esto, está la intuición de que las reglas productivas son las que definen conjuntos infinitos de palabras posibles. Así, según Baayen, la distribución de frecuencias de una categoría morfológica contiene la información necesaria para estimar el número de tipos de una población completa. Este número calculado a partir de los conjuntos de palabras *simplex* es considerablemente menor (*strictly finite*) que el calculado para las reglas productivas, el cual parece aproximarse al infinito (aunque para ciertas categorías no siempre es así[66]). Entonces, esta potencialidad pragmática se refiere a una medida de qué tanto se agota el número pragmáticamente posible de tipos (si es estrictamente finito o tiende al infinito) en cuanto al número total de tipos en el corpus. Este aspecto se mide en términos del cociente del número pragmáticamente posible de tipos (S) sobre el número de tipos en el corpus (V):

$$\mathcal{I}^{(N)} = \frac{S}{V^{(N)}}$$

Nótese que $S = n_0 + V$ (donde n_0 = total de formas que no ocurrieron en el corpus) no se conoce directamente a partir del corpus. Sin embargo, se puede estimar a partir de

[65] *Ibid.* [12], p. 59.

[66] *Ibid.* [12], p. 3.

generalizaciones de leyes estadísticas de datos lingüísticos (leyes de Zipf, Zipf-Mandelbrot y Waring-Herdan-Muller aplicadas en el modelo de Orlov-Chitašvili[67]). De esta manera, Baayen estima los valores S necesarios para calcular la potencialidad pragmática de los sufijos holandeses que considera "rivales", $\mathcal{I}_{-te} = 1.51$ y $\mathcal{I}_{-heid} = 4.43$, lo que le permite concluir que, también mediante este último componente de productividad morfológica, la regla asociada al sufijo *~heid* es más productiva que la del sufijo *~te*, esto es, que el primero goza de una potencialidad pragmática mayor que la del segundo.

El trabajo de Baayen y la familia de enfoques de la fonología de estados finitos ejemplifican la gama de métodos de la morfología computacional. Por un lado, aunque el esquema de estados finitos es el más conocido, al basarse en la formulación de reglas a partir del método introspectivo, no ofrece los elementos que nos permitan descubrir morfemas a partir de un corpus. Por el otro, el método estadístico de Baayen es un trabajo basado en córpora y ofrece definiciones interesantes para medir los diferentes tipos de productividad que se podrían considerar cuantitativamente. Sin embargo, las estimaciones de la productividad de los morfemas presuponen que éstos ya se conocen. Estos trabajos dan por hecho que los morfemas involucrados ya se conocen, cosa que obviamente no siempre se puede asumir. Una investigación empírica previa que permita identificar los morfemas de la lengua estudiada bien podría enriquecerlos.

[67] Por su complejidad no viene al caso describir este procedimiento aquí. Éste se explica con detalle en el capítulo 6 de Baayen, *op. cit.* [12] 1989, pp. 124-187, donde también se describen estas leyes.

1.4 Segmentación automática de palabras

En este apartado se presentan los métodos de segmentación de palabras en morfemas. Es decir, se reseñan los trabajos más sobresalientes de determinación automática de fronteras morfológicas, tanto aquellos que recurren a diccionarios de morfemas para reconocer patrones (*pattern-matching*), como aquellos utilizados en el descubrimiento de morfemas con la menor intervención posible del analista. La gama de métodos incluye el trabajo pionero de Zellig Harris en los años cincuenta, el del ruso N. D. Andreev en los sesenta, el de Josse de Kock en los setenta, los trabajos promovidos por los grupos de Ursula Klenk y Gregor Thurmair en los ochenta, y la aplicación y comparación de varias medidas estadísticas llevada a cabo por Kyo Kageura en los noventa.

Cabe señalar que estos métodos de segmentación automática se presentan como alternativa a los proyectos que determinaban fronteras morfológicas recurriendo simplemente a una costosísisma representación de todo el léxico de las lenguas estudiadas, desaprovechando la poca o mucha regularidad morfológica que pudieran tener.

La primera subsección se ocupa de los sistemas donde el analista codifica su conocimiento de la morfología en alguna estructura de datos como método para *reconocer* los morfemas de la lengua en cuestión. Todas las demás subsecciones presentan los procedimientos de descubrimiento morfológico, es decir, aquellos utilizados para determinar lo más automáticamente posible (con la menor intervención del analista) los morfemas de una lengua a partir de un corpus.

1.4.1 Reconocimiento de patrones

En esta subsección se presentan los procedimientos que reconocen morfemas al compararlos con alguna estructura de información donde se encuentra codificado el *conocimiento* de la morfología de la lengua que se estudie. Se trata de sistemas donde el analista mismo codifica manualmente lo que sabe de una lengua o lo que encuentra acerca de ella en las descripciones gramaticales tradicionales a su disposición.

Aprendizaje morfológico

Gregor Thurmair propuso un esquema que aplicó al alemán y el inglés a mediados de los ochenta. El objetivo de su trabajo fue la segmentación morfológica automática basada en el *aprendizaje* previo de la estructura morfológica. La idea central[68] es poder reducir automáticamente grandes cantidades de palabras (sacadas secuencialmente de textos) a sus formas básicas mediante la ayuda de reglas generadas a partir de una lista de ejemplos proporcionada por el investigador. De esta manera, el término 'aprendizaje' se refiere meramente a la codificación automática de reglas a partir de la información que tenga el lingüista sobre la lengua, es decir, éste último tiene que especificar dónde se segmentan las palabras. Por ejemplo, el analista construye un archivo con una lista de adjetivos alemanes flexionados con marca de dativo como la siguiente:

SCHOENEM - Cut2

REICHEM - Cut2

WILDEM - Cut2

[68]Thurmair, "Ein Morphologisches Prozessorsegment zur Erzeugung von Grundformen mithilfe von Lernverfahren" [132] en Schwarz y Thurmair, eds., *Informationslinguistische Texterschließung* [123]. Georg Olms Verlag, Zürich, 1986, pp. 8-31.

WIRREM - Cut2

⋮

BEQUEM - Cut0

Cada adjetivo de la lista tiene una notación agregada por el lingüista que indica cuántos caracteres de la derecha de cada palabra deben eliminarse para quedarse con la forma básica. En el caso de esta lista el autor del sistema tuvo que agregar la marca 'Cut2' que significa 'quitar los dos últimos caracteres'. Así, para obtener la forma básica del adjetivo con marca de dativo 'SCHOENEM' basta con cortar los dos últimos: 'SCHOEN'. Por otra parte, aquellas palabras como 'BEQUEM' que no sean adjetivos con marca de dativo se asocian (manualmente) a la notación 'Cut0' que le dirá al codificador que no quite ningún caracter. A partir de listas como ésta, el sistema *aprende* a segmentar las palabras y debe *aprender* que en todos los casos, excepto cuando haya una 'U' inmediatamente antes de los dos últimos caracteres (como en 'BEQUEM'), se eliminan las dos últimas letras, incluso cuando se trate de palabras que no ocurrieron en esta lista.

Este tipo de procedimiento, apunta Thurmair, es apropiado sobre todo para lenguas con morfología y sistema de flexión ricos y permite el manejo de las excepciones con la simple inclusión de reglas que rindan cuenta de ellas. Nótese, sin embargo, que no permitir la segmentación de 'BEQUEM' se debe a que se trata de una raíz completa, no de que el dativo se marque de otra manera en esa forma. De esta manera, se puede ver que se trata de sencillas manipulaciones de caracteres que podrían elaborarse un poco más para obtener reglas más motivadas lingüísticamente.

Esto es pertinente sobre todo porque aquí —a la usanza de muchísimos lingüistas com-

putacionales (así como de los generativistas)— las reglas se conciben explícitamente como *hipótesis* sobre los contextos de los grafemas y de cómo se procesan ("Standardhypothesen über den Zusammenhang von Endgraphemen und ihrer Verarbeitung"[69]). procesamiento que muchos quieren entender como cognoscitivo, es decir, del hablante.

Las así llamadas hipótesis (que no son otra cosa que los renglones de listas como la de arriba) se codifican en árboles que luego se utilizan para producir una etiqueta de la categoría lematizada a la que pertenece cada palabra representada por la cadena de caracteres que se analice. En resumen, el sistema tiene dos fases, una para el así llamado aprendizaje y otra de análisis basada en los datos producidos por la primera. Los resultados *erróneos* que se den en el análisis se pueden corregir agregándole a la lista de aprendizaje las *reglas* apropiadas para su manejo correcto.

Codificación de gramáticas

Como se dijo arriba, Thurmair aplicó su procedimiento al alemán y al inglés. Montserrat Meya, por otra parte, ilustra la parte de reconocimiento de patrones del trabajo de Thurmair en su propuesta para el español[70]. La analista utilizó un corpus pequeño del dialecto del español peninsular para comprobar las *hipótesis* que ella misma formuló y que constituyen una gramática de descomposición (*Zerlegungsgrammatik*).

Una diferencia importante entre los procedimientos de Thurmair y de Meya, es que en

---

[69] *Ibid.* [132], p. 9.

[70] Meya, "Morphologische Analyse des Spanischen" [102] en Schwarz y Thurmair, eds., *op. cit.* [123] 1986, pp. 134-156.

el segundo no se construyen árboles con las reglas *hipotéticas* de la morfología española. es decir, no recurre a una fase de aprendizaje (a mi juicio, lo más interesante). En cambio. la gramática de la investigadora describe los tipos de morfema según el tipo de palabra en que aparece (si se trata de un prefijo adverbial o denominal, si es o no la raíz de un verbo, etc.[71]).

El análisis se lleva a cabo mediante una lista de 7,000 morfemas determinados por la analista[72], tanto libres como ligados (raíces y afijos), los cuales permiten que el análisis se lleve a cabo mediante un sencillo proceso de reconocimiento de patrones. Esto requiere de listas de transformaciones de grafías, alomorfos y formas supletivas que rindan cuenta. entre otras cosas, de las modificaciones vocálicas de la flexión, fonemas epentéticos, etc. (por ej.. *feliz* → *felicidad*, *cont~* → *cuent~*, *perd~* → *pierd~*, *produc~* → *produzc~*, etc.). Por último, se toman en cuenta varias propiedades morfológicas específicas a la lengua española. tales como los tres modelos de paradigmas verbales (*~ar*, *~er* e *~ir*) y sus marcas de número. modo. tiempo, aspecto. etc. y los cambios de acentuación en palabras derivadas.

Como se puede ver, en este esquema, como en la mayoría de los estudios automáticos de la morfología, se busca simular varias de las características conocidas de la lengua en cuestión (ejercicio sin duda interesantísimo), pero no de investigarla propiamente a partir de los estados de hecho. Lo mejor de los trabajos basados en el método de Thurmair reside en el proceso llamado de *aprendizaje.* La eliminación de esta fase en el método de Meya. hace que su trabajo no difiera en gran medida de otros trabajos cualitativos. como los presentados

---

[71] *Ibid.* [102], pp. 138-142.

[72] Donde se incluyen todos los morfemas del dialecto peninsular que aparecen en Juilland y Chang Rodríguez, *A Frequency Dictionary of Spanish Words* [73], Mouton, La Haya, 1965, y en la mitad del diccionario de R. Slavý y R. Grossmann, *Wörterbuch der spanischen und deutschen Sprache*, Brandstetter. Wiesbaden, 1975.

a continuación.

Otros métodos cualitativos

Ursula Klenk en varios artículos y dos volúmenes editados por ella[73] presentó diversos trabajos (de diferentes investigadores dedicados a distintas lenguas) que ilustraban la pertinencia de diversos métodos. tanto cualitativos como cuantitativos, en la determinación automática de fronteras morfológicas que no dependiera del listado completo del léxico de la lengua en cuestión.

Klenk misma propuso métodos para el español y el árabe[74]. Para la segunda lengua. por ejemplo, su método se limita al análisis de formas verbales regulares (algunas raíces verbales de tres consonantes). Lo interesante del procedimiento es el carácter discontinuo de la morfología árabe, que se opone al método de estados finitos desarrollado —obviamente— para lenguas de morfología en serie (*anreihender Morphologie*), es decir, que consisten en cadenas de bases y afijos, como las europeas. Así, no es una sorpresa que el esquema de estados finitos sea difícil de aplicar (si acaso es posible, dadas las complicaciones implicadas) a lenguas cuya morfología se manifiesta discontinuamente. Klenk propone un método para especificar las secuencias posibles de vocales y patrones que ella llama esquemas (*Schemata*: "iX*YaZiZ" —forma en imperativo— donde 'X,Y,Z' son las consonantes y '*' indica ausencia de vocal) y que, al descubrir que hacen juego con secuencias particulares de fonemas de las

---

[73]Klenk, ed., *op. cit.* [75. 77] 1992, 1994.

[74]Klenk y Langer, "Morphological Segmentation Without a Lexicon" [79], *Literary and Linguistic Computing*, 4:4 (1989), pp. 247-253; Klenk, "Verfahren morphologischer Segmentierung und die Wortstruktur des Spanischen" [76] en *op. cit.* [75] 1992, pp. 110-124; y Klenk, "Automatische morphologische Analyse arabischer Verbformen" [78] en *op. cit.* [77] 1994, pp. 84-101.

palabras de entrada, sirven para asignar los rasgos morfológicos a estas palabras (83 rasgos complejos, por ej., si se trata de formas pasivas, infinitivas, imperativas, etc.). Así, la palabra de entrada se compara con un diccionario y con estos esquemas para determinar si se trata de una secuencia permitida y, en caso de que así sea, sus rasgos morfológicos. Como puede verse, en esencia se trata de reglas para asignar estructuras morfológicas.

En cuanto a su trabajo con la lengua española. Klenk desarrolló dos programas[75]: MORSPAN y MORGRA. El primero fue trabajo hecho especialmente para el español y el segundo es una extensión que, basada en los mismos principios, se puede aplicar a otras lenguas (también se describe abajo). Klenk simplemente asume la perspectiva tradicional para concebir la palabra española como la secuencia siguiente[76]: afijo de derivación (opcional) + base + afijo de derivación (opcional) + afijo de flexión + clíticos. Su objetivo principal es determinar las fronteras entre la terminación de la palabra (los sufijos de flexión + clíticos) y todo lo que los preceda (base compleja). Todo su método se basa en el hecho de que el segmento final está compuesto generalmente por uno o varios caracteres pertenecientes a un grupo bien definido: r, l, n, s, t, d, b, m, a, á, e, é, i, í, o, ó (se trata de lengua escrita); mientras que la base puede estar formada prácticamente con todas las letras. Esto sirve para determinar un sistema de reglas de segmentación, las cuales se formulan mediante un procedimiento de *descubrimiento*. Pero el descubrimiento no lo hace la máquina, sino el analista, que laboriosamente tiene que examinar cada secuencia de grafemas del corpus para determinar si forman parte de uno de los segmentos de flexión posibles en español.

---

[75]Véanse Klenk y Langer, art. cit. [79] 1989, y Klenk, art. cit. [76] 1992.

[76]Klenk, art. cit. [79] 1989, p. 248.

Frecuencias de combinaciones de letras

Aquí examinaremos un procedimiento cuantitativo para el que las frecuencias de pares de caracteres son indicadores de fronteras morfológicas. Este método ha sido aplicado a varias lenguas (alemán[77], francés[78] y español[79], entre ellas). La idea central del esquema fue descrita por Langer en 1989. Su procedimiento se basa en que hay ciertos pares de letras que ocurren exclusiva o predominantemente en ciertas posiciones definidas morfológicamente. Por ejemplo, pares de caracteres como 'cl', 'cr' y 'qu' tienden a ocurrir (o siempre aparecen) al principio de morfemas, lo que significa que debe haber una frontera morfológica inmediatamente antes de ellos; mientras que 'nm' y 'ks' contienen muy a menudo una frontera morfológica en su interior[80].

Para explotar esta observación. es necesario registrar el porcentaje de ocurrencias de cada combinación posible de caracteres g_1g_2 que ocurre inmediatamente después de una frontera ($A(g_1g_2)$), el de pares que contienen una frontera entre cada componente ($M(g_1g_2)$). aquel de los pares que ocurren seguidos inmediatamente por límites morfológicos ($E(g_1g_2)$) y la proporción de ocurrencias de cada par en contextos que no incluyen ninguno de los tres casos anteriores ($N(g_1g_2)$). Así. la secuencia 'st' en inglés tendrá un porcentaje $A(g_1g_2)$ por su ocurrencia en palabras como 'stand', otro $M(g_1g_2)$ por aparecer en compuestos como 'messtin' (recipiente metálico utilizado por militares), un porcentaje $E(g_1g_2)$ al ocurrir en

---

[77]Programa MOSES en Klenk y Langer, art. cit. [79] 1989, pp. 250-251.

[78]Programa MOSEF en Janßen, "Segmentierung französischer Wortformen in Morphe ohne Verwendung eines Lexikons" [70], en Klenk, ed. *op. cit.* [75] 1992, pp. 74-95.

[79]Programa MOSS en Flenner, "Ein quantitatives Morphsegmentierungssystem für spanische Wortformen" [48] en Klenk, ed. *op. cit.* [77] 1994, pp. 31-62.

[80]Klenk y Langer, art. cit. [79] 1989, p. 250.

vocablos como 'must' y otro $N(g_1g_2)$ por palabras como 'custom'. Con esta información se construye una tabla que especifique cada uno de estos valores para cada posible combinación de caracteres. Luego todo esto se aplica en la segmentación automática de palabras. Véase la tabla 1.5.

Tabla 1.5: Segmentaciones posibles del vocablo 'zerlegen'

g_1g_2	Z[a]	E	R	L	E	G	E	N	
en							89%	1%	85%
ge						63%	34%	61%	
eg					1%	25%	58%		
le				33%	46%	12%			
rl			0%	100%	0%				
er		51%	3%	86%					
ze	51%	49%	11%						
	51	50	5	73	16	33	60	31	85

[a]Tabla basada en la tabla de Langer (en Klenk y Langer, art. cit. [79] 1989, p. 251). El primer valor de cada renglón se refiere a $A(g_1g_2)$, el siguiente a $M(g_1g_2)$ y el último a $E(g_1g_2)$.

Los valores de las columnas se combinan para determinar donde se puede segmentar la palabra. Por simplicidad, Langer saca un promedio. El último renglón contiene estos promedios. Nótese que los dos valores más altos coinciden con las fronteras morfológicas al interior de 'zerlegen': entre el prefijo *zer~*, la raíz *~leg~* y el sufijo de flexión *~en*. Los resultados muestran para el alemán (Klenk y Langer, art. cit. [79] 1989, p. 251) alrededor del 90% de palabras segmentadas correctamente (se eliminaron palabras extranjeras, abreviaciones, etc.), para el francés alrededor de 70% (Janßen, art. cit. [70] 1992, pp. 74, 89-93) y entre 68% y 94% para el español (Flenner, art. cit. [48] 1994, p. 57)[81]. Este es un resultado muy respetable. El único inconveniente es la laboriosísima intervención del analista para determinar los porcentajes de las posiciones de las fronteras en cada par de caracteres, cosa nada trivial.

[81]Los porcentajes de aciertos para las versiones del francés y del español varían según la aplicación de componentes de reglas para mejorar los resultados logrados con este método.

En este grupo de investigaciones, es finalmente también el lingüista quien decide dónde están las fronteras.

Recuperación de Información (*Information Retrieval*)

Entre los métodos más conocidos de segmentación de palabras están los que se han desarrollado en el marco de recuperación de documentos y que se caracterizan por ser programas pequeños y muy ágiles (ocupan poco espacio y son muy rápidos).

El algoritmo de Porter[82] es tal vez el más conocido y, aunque se han hecho versiones para varias lenguas, se diseñó especialmente para el inglés y refleja la sencillez de su morfología. Se trata de un *stemmer* (programa que desnuda las palabras de sus sufijos) y consiste en una secuencia de reglas que de cumplirse ciertas condiciones eliminan de la palabra la cadena de caracteres más larga que pueda recortarse en ese momento, cosa que se repite hasta que ya no se pueden eliminar más caracteres (*iterative longest match stemming*). No hay una estructura representativa de la morfología que esté separada de las instrucciones del algoritmo (se trata de una representación procesal, como se denominó arriba), es decir, la información morfológica está integrada a las reglas condicionales de la manera siguiente:

```
if a word ends in "ies" but not "eies" or "aies"
then "ies" → "y"
if a word ends in "es" but not "aes", "ees" or "oes"
then "es" → "e"
⋮
```

De esta manera, se especifican varias condiciones que la base hipotética debe cumplir para

[82] Véanse Porter, M.F., art. cit. [115] 1980, pp. 130-137; y Frakes, art. cit. [49] 1992, pp. 131-160.

que se elimine de la palabra el sufijo que se cree que se ha encontrado. Las condiciones son las siguientes:

- que contenga un número determinado (`m`) de secuencias VC (donde V puede incluir una o varias vocales y C una o varias consonantes):
  TR, EE, TREE, Y, BY (`m` = 0)
  TROUBLE, OATS, TREES, IVY (`m` = 1)
  TROUBLES, PRIVATE, OATEN (`m` = 2).
- que contenga una vocal: `*v*`,
- que termine con doble consonante: `*d`,
- que termine con cierta letra: `*<X>`, donde `X` = cierta letra.
- que termine en secuencia CVC y que la última C no sea 'w', 'x' o 'y': `*o`.

Como ya se dijo arriba, las reglas especifican que, si se cumple cierta condición en la base y cierto sufijo está involucrado, éste será sustituido por un nuevo segmento (que también puede ser el segmento $\emptyset$, "NULL"). Algunos ejemplos aparecen en la tabla 1.6:

Tabla 1.6: Algunas reglas del algoritmo de Porter

condiciones	sufijo	reemplazo	ejemplos
NULL	sses	ss	caresses → caress
($m > 0$)	ed	NULL	plastered → plaster
			bled → bled
($*v*$)	ing	NULL	motoring → motor
			sing → sing
NULL	s	NULL	cats → cat
($m > 0$)	iveness	ive	decisiveness → decisive
($m > 0$)	alize	al	formalize → formal

Naturalmente el secreto del procedimiento es que las reglas están ordenadas de manera específica. Las reglas se agrupan en pasos y subpasos. El primer paso (constituido por tres subpasos de no más de cinco reglas el más largo) recorta los poquísimos sufijos de flexión del inglés. Los cuatro pasos restantes se ocupan de los sufijos derivativos y contienen un promedio de diez reglas cada uno (el más largo tiene solamente veinte). En esencia aquí no

hay investigación morfológica más allá de la que Porter mismo llevó a cabo revisando sus diccionarios y aplicando su sentido común.

1.4.2 Métodos de estadística de digramas

En esta subsección se presentan las estadísticas de co-ocurrencia de digramas como métodos de descubrimiento de fronteras morfológicas. Estas estadísticas se han aplicado ampliamente tanto en trabajos de extracción de unidades fraseológicas (*collocations*)[83]. como en estudios lingüísticos y literarios basados en córpora.

Un digrama es sencillamente un par de segmentos que ocurren en un corpus, uno después del otro[84]. Hay varias estadísticas para medir la asociación entre los dos elementos de un digrama. Cada una define el concepto de asociación de manera diferente[85]. pero comparten el concepto de *no asociación* que definen en términos de independencia. Es decir. de cada estadística se obtiene un valor que mide la independencia entre segmentos: a menor valor mayor asociación, a mayor valor mayor independencia. La tabla 1.8 resume estas estadísticas. La tabla 1.7 reproduce el tipo de tabla de contingencia a partir de la cual se pueden calcular las estadísticas[86].

---

[83]Como vimos en la página 43, las unidades fraseológicas o 'colocaciones' son expresiones formadas por dos o más palabras y que corresponden a una manera convencional de decir algo.

[84]Un trigrama son tres segmentos, un tetragrama cuatro, etc.

[85]Véanse Manning y Shütze, *op. cit.* [93] 1999, pp. 169-182; y Kyo Kageura (art. cit. [74] 1999, pp. 149-166) para una explicación detallada.

[86]Si en las tablas 1.7 y 1.8 entendemos w como el conjunto de ocurrencias en el corpus de un segmento dado. $\overline{w}$ es su complemento. Así, la frecuencia de $\overline{w}$ se refiere a la frecuencia de todo lo que no es w. pero es un segmento del corpus. De esta manera, mientras $f(w)$ es la frecuencia de w, $f(\overline{w})$ es la frecuencia de todos los segmentos que no son w.

Como antes se mencionó, la prueba de independencia de χ^2 que es un método muy socorrido, principalmente porque no es necesario asumir que los datos se distribuyen normalmente. En esencia, se trata de comparar las frecuencias observadas con las frecuencias que cabe esperar si las poblaciones contrastadas fueran independientes la una de la otra. La estadística se calcula mediante la siguiente fórmula[87]:

$$X^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

donde O representa las frecuencias observadas y E representa las esperadas[88]. De esta fórmula se deriva la de la tabla 1.8. Así, si la diferencia entre lo observado y lo esperado es lo suficientemente grande, entonces se presume independencia entre las dos poblaciones. El problema principal de esta prueba es, como ya se dijo, que no es apropiada para muestras pequeñas o cuando alguna de las celdas de la tabla tiene un número menor a cinco.

Tabla 1.7: Tabla de contingencia para el digrama $w_1 w_2$.

	w_2	$\overline{w_2}$	**total**
w_1	$f(w_1 w_2)$	$f(w_1 \overline{w_2})$	$f(w_1)$
$\overline{w_1}$	$f(\overline{w_1} w_2)$	$f(\overline{w_1} \overline{w_2})$	$f(\overline{w_1})$
total	$f(w_2)$	$f(\overline{w_2})$	$T = \sum f(w_i)$

La razón de semejanza constituye otro método para determinar si hay asociación entre dos poblaciones, que es más apropiado para muestras escasas que la prueba de χ^2. En esencia, se trata de determinar si la probabilidad de un evento A dado un evento B es igual a la probabilidad de que ocurra A cuando B no ha ocurrido:

$$P(A|B) = P(A|\overline{B})$$

[87]Convencionalmente, χ^2 se reserva para la distribución, por lo que se usa X^2 para la estadística.

[88]Las frecuencias esperadas se calculan mediante los totales —en la tabla 1.7— de los renglones y las columnas convertidos a proporciones del total T.

Tabla 1.8: Estadísticas para medir no asociación entre digramas.

estadística[a]	definición de las medidas[b]
χ^2	$X^2 = \frac{T((f(w_1w_2)f(\overline{w_1}\overline{w_2}))-(f(\overline{w_1}w_2)f(w_1\overline{w_2})))^2}{f(w_1)f(\overline{w_1})f(w_2)f(\overline{w_2})}$
razón de semejanza	$-2log\lambda =$ $2\left[\left(log(L(\frac{f(w_1w_2)}{f(w_2)}, f(w_1w_2), f(w_2))) + log(L(\frac{f(w_1\overline{w_2})}{f(\overline{w_2})}, f(w_1\overline{w_2}), f(\overline{w_2}))\right)\right.$ $\left. - \left(log(L(\frac{f(w_1)}{T}, f(w_1w_2), f(w_2))) + log(L(\frac{f(w_1)}{T}, f(w_1\overline{w_2}), f(\overline{w_2})))\right)\right]$. donde $L(p,n,k) = nlog(p) + (k-n)log(1-p)$
coef. coligación de Yule	$Y = \frac{\sqrt{a}-1}{\sqrt{a}+1}$, donde $a = \frac{f(w_1w_2)f(\overline{w_1}\overline{w_2})}{f(w_1\overline{w_2})f(\overline{w_1}w_2)}$
información mutua	$I = log_2\left(\frac{f(w_1w_2)/T}{(f(w_2)/T)(f(w_1)/T)}\right)$

[a] χ^2 se reserva para la distribución, por lo que se usa X^2 para la estadística. Esta fórmula no toma en cuenta ningún ajuste para datos escasos.

[b] Las frecuencias se refieren a la tabla 1.7.

para lo que se asume una distribución binomial. De esto se deriva la larga fórmula[89] de la tabla 1.8.

El coeficiente de coligación de Yule se basa en una medida de asociación comúnmente conocida como producto cruzado (*cross-product ratio* u *odds-ratio*)[90]: $a = \frac{f(w_1w_2)f(\overline{w_1}\overline{w_2})}{f(w_1\overline{w_2})f(\overline{w_1}w_2)}$. En esencia se trata de un índice que compara los dos renglones de la tabla 1.7.

La estadística llamada de información mutua específica (*specific or pointwise*) se refiere a la información (medida mediante el logaritmo de base 2) entre dos eventos particulares (para nuestros objetivos, dos segmentos de palabras). Se define mediante la siguiente fórmula :

$$log_2\frac{P(AB)}{P(A)P(B)} = log_2\frac{P(A|B)}{P(A)} = log_2\frac{P(B|A)}{P(B)}$$

[89] Para explicación detallada, véase Manning y Schütze, *op. cit.* [93] 1999, pp. 172-173.

[90] Véase Kageura art. cit. [74] 1999.

La idea es medir qué tanta información (en el sentido técnico, que se describirá más adelante, a partir de la página 83) nos proporciona el que una palabra ocurra en relación con la ocurrencia de la otra. Esta fórmula corresponde a la última de la tabla 1.8.

Todas estas medidas significan cosas distintas, sobre todo si se trata de determinar la asociación de dos elementos (como ya se dijo miden diferentes tipos de asociación). Sin embargo, cuando se trata de determinar la no asociación resultan relativamente comparables, porque al no haber asociación, no importa qué tipo de asociación se esté buscando.

En el experimento de Kageura se comparan estas medidas para determinar cuál es más apropiada en la determinación de fronteras morfológicas en secuencias de caracteres Kanji del japonés. En su experimento la razón de semejanza resultó ser la mejor medida, seguida por la prueba de χ^2. Estos resultados, como se verá más adelante, no coinciden con los obtenidos en este capítulo para el español (véase la página 120).

1.4.3 Frecuencias de caracteres (la escuela rusa)

En esta subsección se comenta el trabajo del equipo ruso dirigido por N. D. Andreev que trabajó hace casi cuarenta años en el reconocimiento de afijos de diversas lenguas. Este es el primer procedimiento automatizado de descubrimiento morfológico. Data de la primera mitad de los años sesenta y se llevó a cabo en la entonces Unión Soviética. Andreev y su equipo desarrollaron y aplicaron al ruso, al húngaro, al vietnamita y a otras lenguas un programa capaz de determinar empíricamente afijos de flexión. Oliver Cromm[91] construyó

[91] Cromm, *op. cit.* [46] 1996.

en 1996 un programa para aplicar el método de Andreev al alemán.

El procedimiento de Andreev está basado en el hecho de que los afijos en lengua escrita son cadenas de caracteres que se caracterizan por ser estadísticamente mucho más frecuentes que otras cadenas de caracteres (otros tipos de morfemas) y en que tienden a aparecer sistemáticamente con las mismas bases con que aparecen otros afijos. La idea es encontrar mediante la comparación de muestras de segmentos de palabras y sus frecuencias. así como mediante la búsqueda de combinaciones de segmentos[92]. los paradigmas en que ocurren los afijos investigados. Pero en lugar de contar fonemas anteriores y posteriores como lo hace Harris (véase más adelante), Andreev utiliza frecuencias de caracteres según sus posiciones dentro de la palabra.

El método consiste en buscar en los extremos de las palabras (entre las primeras y las últimas letras) las cadenas de caracteres más frecuentes en esa posición en la totalidad de palabras del texto examinado. La idea es que los afijos son más frecuentes que las bases y eso se debe reflejar en la observación de estas frecuencias. Las secuencias más frecuentes se examinan luego para determinar si son o no afijos. Las secuencias restantes (aquellas que quedan después de eliminar los presuntos afijos de las palabras) se consideran entonces posibles bases. Estas últimas se examinan para determinar si aparecen combinadas con otros segmentos que se intercambian (según la evidencia en el texto examinado) con el primer presunto afijo. Si hay varios candidatos a bases que se combinan con varios presuntos afijos que se intercambian entre sí, se ha encontrado un paradigma.

Con el objeto de llevar todo esto a cabo, se calculan varios parámetros para determinar

[92]Que Greenberg llamara "cuadros" (*squares*); véase definición de *cuadro* más adelante en la página 104.

qué combinaciones se deben menos al azar y para distinguir los paradigmas de flexión de los de derivación, siendo los primeros el objetivo principal del procedimiento. Los índices principales[93] son los siguientes: medida de desnivel (en ruso *mera perepada* o en alemán *Gefällemaß*), función correlativa (ruso *korrelativnaja funkcija*. alemán *korrelative Funktion*) y medida de reducción (*mera redukcii* o *Reduktionsmaß*).

Para calcular la primera medida es necesario obtener las frecuencias de los caracteres y luego ordenarlas de mayor a menor: $f_{x_1} \geq f_{x_2} \geq \ldots f_{x_n}$ (donde n es el número de caracteres y x_1 es el caracter más frecuente, x_2 el próximo más frecuente, ... y x_n el de menos ocurrencias). Al ordenarlas por probabilidad (frecuencia relativa) se obtiene el mismo orden: $P_{x_1} \geq P_{x_2} \geq \ldots P_{x_n}$. La medida de desnivel es simplemente la probabilidad del objeto x_i $(1 \leq i \leq n)$ con respecto a la del siguiente más frecuente:

$$\frac{P_{x_i}}{P_{x_{i+1}}} \geq 1$$

Cromm aplicó el procedimiento de Andreev a una versión electrónica de la biblia en alemán. En la tabla 1.9 aparecen ordenados los caracteres más frecuentes según su probabilidad relativa absoluta. La última columna contiene la medida de desnivel.

La segunda medida, la función correlativa de un caracter o cadena de caracteres (al principio o al final de una palabra), es su frecuencia relativa en esa posición (la probabilidad de que ocurra allí) dividida entre la probabilidad de ocurrencia de dicho objeto en cualquier posición del corpus. Es decir, el cociente de la probabilidad de que un objeto x ocurra en la

[93] Cromm *op. cit.* [46] 1996, p. 23-26.

Tabla 1.9: Frecuencias de caracteres de la biblia en alemán

car.	fr.	fr. rel.	Gefällemaß[a]
e	584376	0.145	1.667
n	348073	0.087	1.381
r	252659	0.063	1.086
i	128328	0.058	1.094
h	213431	0.053	1.000
a	211093	0.053	1.060
t	199464	0.050	1.087
s	183680	0.046	1.438
⋮	⋮	⋮	⋮
q	155	0.000	—
x	40	0.000	—

[a]Se trata aquí de frecuencias absolutas. La misma medida de desnivel se calcula para cada posición de los caracteres en la palabra y *Gefällemaß* ≥ 1.5 es el umbral considerado decisivo en la presunción de una frontera de morfema.

posición j, entre su probabilidad de ocurrir en cualquier parte:

$$KF = \frac{P_x^j}{P_x}$$

Tabla 1.10: Función correlativa KF de los caracteres más frecuentes según su posición en la cadena de caracteres

car.	-4	-3	-2	-1	1	2	3	4
e	1.0	0.9	2.6	1.1	0.4	2.0	0.6	1.2
n	0.7	1.0	1.1	3.3	0.5		1.4	1.3
r	0.9	1.1	1.1	1.6	0.3	0.9	2.5	1.2
t	0.8	1.6	1.1	2.9	0.4		1.1	1.5
j					7.1			
v					6.5			
w					5.6			
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

El objeto (caracter o cadena de caracteres) que en cierta posición j obtenga el valor KF más alto es denominado informante (*Informant*). Por ejemplo, en la tabla 1.10 se ve que el informante de la última posición (-1) de las palabras alemanas es el caracter 'n', cosa que indica su posible uso como frontera de flexión (lo cual es cierto en esa lengua). En la penúltima posición (-2) el informante es 'e', lo que indica que por lo menos existe un sufijo de dos caracteres que empieza con este caracter (en alemán, de hecho, hay varios: *~en*, *~er*,

~*es*. etc.).

El algoritmo tiene pasos muy específicos que no es necesario considerar aquí, pero que se basan en estos criterios de frecuencia para determinar afijos y los paradigmas de flexión a los que pertenecen. También con ese mismo objetivo se calculan D (longitud promedio de las palabras en número de letras) y F (tamaño promedio de oración o cantidad promedio de palabras en una oración). V es el número de palabras en el corpus. El procedimiento sólo toma en cuenta vocablos de más de $\frac{2}{3}D$ letras (dos tercios del promedio de caracteres por palabra)[94] o más de cuatro caracteres de longitud.

Entonces, el algoritmo recorre cada vocablo examinando cada una de sus posiciones (del exterior al interior) y calcula en cada una la función correlativa KF (sólo para aquellas letras con mayor frecuencia en esa posición), luego estima las probabilidades de las letras pertinentes (para obtener una medida de desnivel para esa posición) y las compara para elegir los candidatos a afijos.

En la secuencia de caracteres del vocablo analizado cada afijo hipotético tiene un número determinado de bases M_i (donde i es el número del afijo al que se le asocian M_i bases). Eso significa que para el afijo 1 (el más numeroso), hay un número M_1 de bases y para el 2 (que contiene al segmento del hipotético afijo 1), un número M_2 de posibles bases. Entonces, la medida de reducción se obtiene con la siguiente fórmula:

$$\frac{M_1 - M_2}{M_1 K}, \text{ donde } \log_{10} K = \frac{\log_{10} D}{1 + 0.02 \log_{10} V}$$

[94]Por qué $\frac{2}{3}D$ o por qué otros valores que se utilizan en el algoritmo (algunos de los cuales aquí no es necesario presentar) se definen como se definen son preguntas pertinentes. Estas medidas pueden parecer un tanto arbitrarias pero reflejan decisiones informadas (sobre la relación entre D, F y V en distintas lenguas) que ya se conocían o surgieron en el trabajo empírico de ensayo y corrección.

donde K es un coeficiente de reducción. Esta medida[95] sirve para juzgar si se trata o no de afijos de flexión: Mientras más grande sea el número de bases asociado al afijo 2, menor será esta medida de reducción y más seguro será decidir que este afijo es flexivo y no derivativo (y menos probabilidad habrá de que se trate de un error).

Por último, las supuestas bases que resultan de este procedimiento también se examinan cuantitativamente (de hecho, hay todavía otras medidas, por ej. de descentralización, *Dezentrationsmaß*, que aquí no es necesario describir) con el objeto, como se dijo arriba, de detectar los paradigmas de los que los afijos forman parte.

1.4.4 Cuentas de fonemas anteriores y posteriores

En esta subsección se presenta el procedimiento propuesto en los años cincuenta por Zellig Harris para segmentar palabras en morfemas. En un artículo muy leído[96], Harris exploró la correspondencia de fronteras morfológicas y el número de signos que potencialmente siguen o preceden a algún fonema en una expresión dada. Su método consistía en que, dada una segmentación de palabra, se comparaba la variedad de fonemas sucesores y predecesores potenciales. Mientras más grande fuera la variedad de fonemas que potencialmente aparecen antes o después de una segmentación, mayor es la incertidumbre de lo que sigue o precede: lo cual significa que hay una gran posibilidad de tener allí una frontera morfológica.

El procedimiento requiere un corpus de buen tamaño, de preferencia un conjunto de enunciados obtenidos (elicitados) de algún informante. Dado que se trata de una investigación

[95] En su experimento Cromm la redujo a $\frac{M_2}{M_1}$.

[96] Harris, art. cit. [65] 1955, pp. 190-222.

morfológica (y no de la distribución de las letras utilizadas en la escritura de la lengua estudiada), todos los enunciados deben escribirse mediante la misma representación fonética (que no haga referencia, por supuesto, a la representación de morfemas). La mayoría de las segmentaciones propuestas mediante este método coincide con las fronteras entre palabras y morfemas al interior de éstas. Por ejemplo, mediante este procedimiento, */hiyskwik•r/* ('he's quicker') se segmenta */hiy.s.kwik.•r/*. Nótese que no hay nada que indique el estatus morfológico de los segmentos, es decir, que se trate de palabras o pedazos de palabras (no hay nada que indique si /•r/ es sufijo o palabra). Por eso, las decisiones en cuanto al estatus morfológico se le dejan a los otros métodos distribucionales de determinación de morfemas. Es decir, este procedimiento es solamente un intento de poner orden a aquellos anteriormente propuestos en el marco del distribucionalismo[97]. El primer paso es determinar el número de fonemas que sigue a cada segmentación examinada. Por ejemplo, al examinar la segmentación entre el primer fonema común a varios enunciados y los fonemas que le siguen en esos enunciados, simplemente se cuenta el número de fonemas diferentes que ocurren en la segunda posición de esos enunciados. Luego se toman aquellos enunciados que empiezan con una secuencia común de dos fonemas y se cuenta el número de fonemas diferentes que ocurren en la tercera posición en esos enunciados. Así, este procedimiento se repite hasta alcanzar el final de un enunciado particular. En la secuencia de cuentas de fonemas obtenidas para un enunciado de esta manera, se puede ver cómo los números más altos constituyen picos rodeados de hendiduras (la secuencia sube y baja). En cada pico se presume la existencia de una segmentación morfológica. A este procedimiento base se le pueden hacer ciertas modificaciones para afinar los resultados. La más importante consiste en contar no nada

[97] Harris, art. cit. [65] 1955, p. 191.

más los fonemas que siguen, sino también los que preceden la segmentación examinada. Lo sorprendente es que los picos de estas cuentas también corresponden a cortes morfológicos en el enunciado. Nótese la coincidencia entre las fronteras morfológicas y las cuentas altas en ambas direcciones del enunciado inglés que aparece en la tabla 1.11. Las cuentas posteriores

Tabla 1.11: Cuentas de fonemas anteriores y posteriores en cada segmentación de *What did he think of?*

cuentas[a]	**h**	**w**	**•**	**t**	**d**	**ɪ**	**d**	**h**	**i**	**y**	**θ**	**i**	**n**	**k**	**ə**	**v**
posteriores	9	5	1	29	10	19	28	8	12	28	5	4	1	29	11	28
anteriores	22	1	7	18	23	1	3	9	19	4	22	15	3	12	23	6

[a]Tabla basada en la de Harris, art. cit. [65] 1955, p. 218. • = schwa.

muestran picos que coinciden con la segmentación morfológica (inmediatamente después del pico). Lo mismo las anteriores (con excepción entre */dɪd/* y */hiy/*) que muestran los picos antes de la segmentación.

Otra modificación es una operación de inserción que consiste simplemente en meter entre el fonema n y el fonema $n+1$ de un enunciado alguna secuencia fonémica, de tal manera que el resultado sea uno de los enunciados que se atestiguan en el corpus[98]. El número de enunciados atestiguados en el corpus que se puedan obtener mediante esta operación se cuenta como otra medida de la validez de la segmentación. Otra modificación[99] consiste en tomar en cuenta no nada más la variedad de fonemas adyacentes a la segmentación examinada ($n+1$), sino también aquella de los fonemas que se encuentran a dos posiciones de distancia ($n+2$). La cuenta de la variedad de estos fonemas es también una medida de incertidumbre (lo poco predecible) de lo que hay al otro lado de la frontera morfológica: de nuevo, a mayor variedad de lo esperado, mayor certeza de que la segmentación examinada es morfológica. La última

[98]Harris, art. cit. [65] 1955, p. 199.

[99]Harris, art. cit. [65] 1955, pp. 199-202.

modificación consiste en tomar en cuenta los tipos de fonemas anteriores y posteriores. La idea es que hay tipos de variedades más probables después (y antes) de ciertas secuencias. Por ejemplo, después de una secuencia que termina en consonante, hay una gran probabilidad de que la variedad posible de fonemas sea predominantemente vocálica. Para corregir fluctuaciones debidas a las diferencias en número entre los conjuntos de vocales y consonantes. Harris propone fórmulas que reflejan la fonotáctica particular de la lengua examinada.

1.4.5 Teoría de la información

Muy ligado al método de Harris está la aplicación del concepto de entropía, que con frecuencia, como ya se dijo en la introducción, se menciona como un método apropiado para determinar morfemas. En esta subsección se presentan los fundamentos del procedimiento.

La teoría de la información (también conocida como teoría de la comunicación)[100] es una teoría matemática que permite analizar cuantitativamente la incertidumbre. En esta teoría, 'incertidumbre' equivale a 'información' (que aquí es una noción técnica que no es idéntica a la noción intuitiva de información en cuanto al significado se refiere[101]). La teoría de la información se ocupa de la *cantidad* de información que lleva una señal y no del contenido de dicha señal.

La idea detrás de asociar incertidumbre con información es que cierta cantidad de infor-

[100] Fue desarrollada a finales de los años cuarenta por Claude Shannon y Warren Weaver, *op. cit.* [125] 1964; véanse Meyer-Eppler, *Grundlagen und Anwendungen der Informationstheorie* [103], Springer, Heidelberg, 1969 y Fernández García, *Acerca de la teoría de la información y algunas de sus aplicaciones* [47], Departamento de Matemáticas, Facultad de Ciencias, UNAM, 1978.

[101] Nótese que una cadena aleatoria de signos genera más información que una de signos que constituyen un texto con algún sentido, ya que la segunda cadena es más predecible o menos incierta.

mación es necesaria para identificar correctamente un mensaje a partir de varios mensajes posibles (resolver la incertidumbre), es decir, cuando a partir de una señal recibida uno tiene que decidir de entre un conjunto de mensajes posibles, cuál es el correcto. Así, medir la incertidumbre equivale a establecer la cantidad de información necesaria para resolverla (para hacer cierto lo incierto). También se utilizan términos como 'sorpresa' (porque cualquier medida de información causa una cantidad proporcional de sorpresa) y con mucha frecuencia 'entropía' (caos, desorden o energía desorganizada), debido a la relación obvia con este concepto, originado en la termodinámica.

La entropía o incertidumbre[102] de una variable discreta X —que toma un número finito de valores $x_1, x_2, x_3, ..., x_n$, cada uno con una probabilidad $p_1, p_2, p_3, ..., p_n$, donde $0 \leq p_i \leq 1$ $(i = 1, 2, 3, ..., n)$ y la suma de las probabilidades es igual a 1 $(\sum_{i=1}^{n} p_i = 1)$— se calcula mediante la siguiente fórmula:

$$H(X) = H(p_1, p_2, p_3, ..., p_n) = -\sum_{i=1}^{n} p_i \times log_2(p_i) \qquad (1.1)$$

donde $p_i \times log_2(p_i) = 0$, si $p_i = 0$.

Supongamos que tenemos 2 mensajes o símbolos posibles $(X = \{x_1, x_2\})$. La figura 1.2 muestra la relación entre la entropía o incertidumbre y la probabilidad de uno de esos eventos (p_1); la probabilidad del otro sería obviamente $p_2 = 1 - p_1$. Se puede apreciar que la entropía es mayor cuando la probabilidad está repartida equitativamente entre los dos valores (es decir, cuando son equiprobables: $p_1 = p_2 = 0.5$), mientras que disminuirá cuando alguna de las dos probabilidades se acerque a cero (por lo que lo otra se acercaría a uno). En otras palabras, la incertidumbre disminuiría cuando la probabilidad de alguno de los dos eventos se

[102]Véanse Shannon *op. cit.* [125] 1964, p. 50; Weaver art. cit. [134] 1964, p. 14.

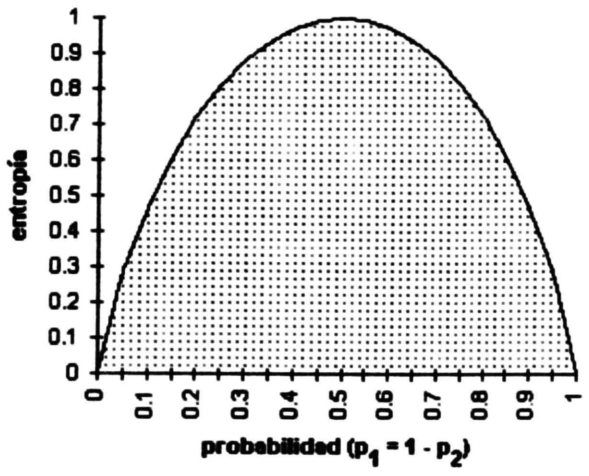

Figura 1.2: Gráfica de la entropía de dos mensajes posibles $X = \{x_1, x_2\}$

acerque a cero y, cuando las probabilidades sean equivalentes, la incertidumbre de obtener tal o cual mensaje sería la mayor posible[103] (mayor sería la sorpresa si predecimos correctamente el próximo símbolo).

La unidad de la entropía calculada mediante el logaritmo de base dos se llama *bit* (contracción de BI*nary digi*T). Ya que la función en la figura 1.2 fue calculada con el logaritmo de base dos y se trata de un sistema de dos probabilidades, la entropía representada allí no puede tener un valor mayor que uno (1 bit), cosa no necesariamente verdadera para sistemas de más de dos probabilidades. Si se calcula la entropía con el logaritmo de otra base, como por ejemplo de 10, se habla de "unidades decimales" de la entropía[104].

El método para descubrir morfemas consiste en medir la entropía de cada segmentación de cada vocablo, es decir entre los dos segmentos del vocablo dividido por esa segmentación.

[103] Weaver art. cit. [134] 1964, p. 15.

[104] Fernández, *op. cit.* [47] 1978, p. 8.

La idea es determinar la frecuencia de todo lo que en el corpus está atestiguado como acompañante de uno de los segmentos y calcular las probabilidades de cada objeto posible después del segmento dado. En la figura 1.3 se ilustra esto. La entropía se obtiene al aplicar

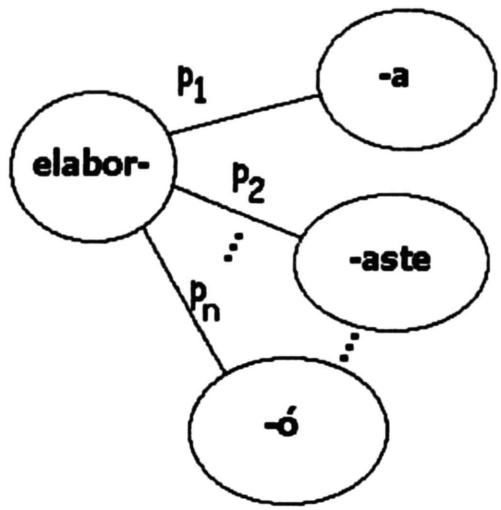

Figura 1.3: Esquema para ilustrar las probabilidades de los segmentos que según un corpus ocurren después de *elabor~*.

la fórmula 1.1 a estas probabilidades. Como veremos más adelante, las segmentaciones que exhiban las cantidades mayores de entropía son el inicio de una raíz porque son los segmentos más informativos del discurso, mientras que los afijos, si bien también informativos, llevan solamente la información estructural del discurso. Este es uno de los métodos aplicados al *CEMC* en esta investigación, así que los mejores ejemplos para demostrar su aplicación se presentan más adelante (en el capítulo siguiente, en la sección sobre entropía que empieza en la página 106).

1.4.6 El principio de economía

En esta sección se presenta el trabajo que Josse de Kock y Walter Bossaert desarrollaron en los años setenta para segmentar palabras del francés y del español. Se trata de otro método para determinar empíricamente las fronteras morfológicas entre bases y afijos, tanto de flexión como de derivación. El método coincide en varios aspectos con el de Andreev y, aunque no llega a ocuparse en la determinación de paradigmas, sienta las bases que permiten hacerlo.

La base de todo el procedimiento es el principio de economía de signos o rentabilidad del sistema. Como es bien sabido, este principio tiene diferentes aspectos, pero el aquí pertinente puede parafrasearse de la siguiente manera: el número de signos en todos los niveles del lenguaje debe ser menor al número de cosas nombradas; así, "the code is organized in such a way that a sign can serve in more than one instance without creating any ambiguity"[105]. De esta manera, los signos del nivel sintáctico de una lengua como el español son producto de la combinación de los signos del nivel morfológico (se flexionan o derivan), siendo estos últimos pocos en número, pero más frecuentes que los primeros. El que la concatenación de dos signos de un nivel sea económica viene a ser entonces una función de esta diferencia, es decir, mientras menos signos de más frecuencia existan en el nivel morfológico que den lugar a más signos (de baja frecuencia) del nivel sintáctico, la lengua será más económica. Y si suponemos que las lenguas tienden a la economía de signos, esta sencilla diferencia de carácter formal nos proporciona un mecanismo para determinar los signos morfológicos.

[105] Josse de Kock y Walter Bossaert, *op. cit.* [82] 1978, p. 15; *op. cit.* [80] 1974: Previamente al experimento de este capítulo, llevé a cabo otro en donde se aplica el método de de Kock y Bossaert a la nomenclatura del DEM, "Un experimento cuantitativo de determinación de fronteras morfológicas del español de México" [96], IV Congreso Internacional de Lingüística en el Noroeste, Hermosillo, Sonora, noviembre 1996.

Dicho de otra manera y en términos de los procesos de diversificación y unificación[106], si el sistema necesita de un número reducido de signos para ser económico, (porque un número pequeño beneficia al hablante de una lengua, al requerirle un menor esfuerzo en el ejercicio de la memoria), estos signos se pueden combinar en otro nivel para dar lugar a un inventario mucho más grande (en beneficio del oyente, al existir mayor diversificación de estructuras que le aclaren el mensaje).

De Kock y Bossaert se basaron en los diccionarios de frecuencias de Alphonse Juilland[107]. En su procedimiento, los investigadores examinan automáticamente cada segmentación posible de cada vocablo, según se encuentren segmentos que aparezcan en varios otros vocablos. Así, el número de vocablos con un segmento *común* a la izquierda, es decir, el número de segmentos *diferentes* que aparecen a la derecha es m_d (*droit*). El número de vocablos con un segmento común a la derecha, o lo que es lo mismo, el número de segmentos distintos a la izquierda es m_g (*gauche*)[108]. También, como en el método de Andreev, calculan para cada segmentación un número de combinaciones de cuatro segmentos[109]. El número de combinaciones que se puedan determinar para cada segmentación se llama n_c (*carré*) y es una medida

Tabla 1.12: Hipótesis para cada corte de los vocablos 'capacidad' y 'olvidad'

vocablo[a]	v b,g	÷ v a,d	= v b,a	v b,d	÷ v a,g	= v a,b
kapaθida :: *d*	0	0		1	17	0.059
kapaθid :: *ad*	0	0		9	6	0.167
kapaθi :: *dad*	0	0		1	16	0.063
kapaθ :: *idad*	36	2	18.000	2	51	0.039
kapa :: *θidad*	1	2	0.500	2	3	0.667
kap :: *aθidad*	0	0		2	1	2.000
ka :: *paθidad*	0	0		0	0	
k :: *apaθidad*	0	0		0	0	
olbida :: *d*	48	11	4.364	9	16	0.563
olbid :: *ad*	29	7	4.143	22	25	0.880
olbi :: *dad*	0	0		21	6	3.500
olb :: *idad*	0	0		4	1	4.000
ol :: *bidad*	0	0		0	0	
o :: *lbidad*	0	0		0	0	

[a]Tabla basada en la de Kock y Bossaert, *op. cit.* [82] 1978, p. 31; **v b,g** = número de segmentos distintos (que aparecen en cuadros) de la izquierda bajo la hipótesis de que la base está a la izquierda; **v a,d** = número de segmentos distintos de la derecha bajo la hipótesis de que el segmento de la derecha es un afijo; **v b,a** = valor de la hipótesis de que el segmento de la izquierda sea la base y el de la derecha un afijo, etc.

de la validez de la segmentación.

El mecanismo más importante es el examen de dos hipótesis para cada corte: ya sea que el segmento a la izquierda sea un prefijo y el de la derecha la raíz, o que el de la derecha sea un sufijo y el de la izquierda una raíz. Así, se calculan dos valores de segmentación: uno dividiendo el número de segmentos a la izquierda propuestos como raíz entre el número de segmentos a la derecha propuestos como afijo, el otro dividiendo los segmentos a la derecha propuestos como raíz entre los de la izquierda propuestos como afijo. El segmento más

[106]Para una descripción general de relación de estos procesos con el tamaño del inventario del léxico de las lenguas, véase Köhler, "Diversification of Coding Methods in Grammar" [83], en Ursula Rothe, ed., *Diversification Processes in Language* [121], Rottmann, Hagen, 1991.

[107]Para el francés, Juilland, *Frequency Dictionary of French Words*, Gembloux, 1965; para el de español, Juilland y Chang Rodríguez, *op. cit.* [73] 1965.

[108]Estos números se modifican al eliminar ciertos segmentos y tomar en cuenta ciertos criterios como que los segmentos contengan vocales o no, o que haya fonemas en un segmento que pertenezcan con más probabilidad al otro. Los números modificados se representan mediante M_d y M_g (de Kock y Bossaert, *op. cit.* [82] 1978, pp. 22-23).

[109]Los mismos que, como se dijo arriba, Greenberg llamara "cuadros" (*squares*); véase definición de *cuadro* más adelante en la página 104.

probable como raíz será aquél cuyo valor calculado bajo la hipótesis de que es la raíz sea mayor a uno. Así, de Kock y Bossaert muestran los ejemplos que aparecen en la tabla 1.12. En el capítulo siguiente, (a partir de la página 111) y mediante una formalización utilizando conjuntos, se parafrasea con detalle el procedimiento de de Kock y Bossaert y su aplicación en el experimento del capítulo sobre el afijo.

1.5 Observaciones finales

En este primer capítulo se presentó un breve panorama general de los trabajos y enfoques que suelen aplicarse al estudio del lenguaje mediante computadoras. Primero se hizo un brevísimo recuento de los métodos dedicados al exterior de las palabras, es decir, sintácticos. Específicamente se mencionan los formalismos más conocidos (unos con pretensiones lingüísticas y otros no, pero todos con el propósito de servir de marco para la generación y síntesis de lenguajes naturales) y sus fundamentos en la escuela generativa (como la tipología de lenguajes de Chomsky, entre otras cosas). Luego, no exactamente en el nivel sintáctico, sino más bien en el léxico, se presentaron los métodos para medir la asociación entre palabras gráficas que pueden aplicarse al descubrimiento de unidades fraseológicas o, de más interés para este trabajo, de palabras gramaticales. Finalmente, en la parte dedicada al nivel morfológico se revisaron algunos enfoques y métodos de trabajo muy conocidos en el ámbito de la morfología automática y, en una sección aparte, los procedimientos para la segmentación morfológica de palabras.

Capítulo 2

El afijo en el *CEMC*

Las metas de este capítulo corresponden al primer y segundo objetivos de la tesis, que son determinar empíricamente un conjunto de signos del nivel morfológico y, a partir de éstos, construir un segmentador de palabras que separe las raíces de los afijos[1]. En el capítulo anterior se presentaron los enfoques más conocidos de la morfología automática, inclusive los métodos de descubrimiento automático de morfemas. En este capítulo, se empieza por definir las nociones de —entre otras— morfema, morfo y afijo. En seguida se abordan algunas cuestiones de carácter formal. Luego se analizan y comparan diferentes técnicas de segmentación de palabras con el objeto de elaborar un método para construir uno o varios catálogos de afijos a partir del *CEMC*. Entonces, a partir de estas reflexiones se examina y matiza la hipótesis de afijalidad como criterio para segmentar palabras. Al final del capítulo se resumen los resultados mediante la presentación de los afijos más importantes determinados automáticamente.

---

[1] Una versión preliminar de este capítulo aparece en Medina Urrea, "Automatic Discovery of Affixes by Means of a Corpus: A Catalog of Spanish Affixes" [97], *Journal of Quantitative Linguistics*, 7:2 (2000). pp. 97-114.

2.1 Sobre las unidades morfológicas

En esta sección se delimitan las nociones morfológicas pertinentes al experimento de este capítulo y se establece el tipo de unidad apropiado para la construcción de un catálogo de signos mínimos para este proyecto.

Como es bien sabido el *morfema* puede entenderse como una abstracción que contiene una o varias formas mínimas que comparten un contenido y que Eugene Nida llamó *alomorfos*[2] (las ocurrencias del morfema en contextos determinados). A grandes rasgos, este concepto de morfema implica que una misma unidad significativa (de la lengua) puede exhibir más de una forma, según el lugar donde ocurra dicha forma (en el habla). Es decir, varias formas pueden compartir un significado: uno o más alomorfos constituyen un morfema. Así, el morfema para construir el {plural de sustantivos} en español puede manifestarse mediante cuando menos los alomorfos ~s, ~es y el morfema ~idad mediante ~edad (variedad), ~dad (beldad), etc., la distribución de los cuales depende de la base a la que se adhieren. No es necesario decir que la investigación automática de los morfemas así definidos conlleva dificultades que por ahora tendrán que resolverse con la intervención del analista, que manualmente decidirá qué formas son ejemplos de qué morfemas.

Es obvio que también es posible que varios morfemas compartan un alomorfo, es decir, que existan signos mínimos con varios significados[3]. Estos signos mínimos, conocidos como

---

[2]Véase Nida, "The Identification of Morphemes" [109], *Language* 24, p. 420. Zellig Harris ya antes los había llamado 'alternantes de morfema', véase, "Morpheme Alternants in Linguistic Analysis" [64]. *Language* 18, p. 170. Compárense las definiciones de morfema, alomorfo y morfo en Glück, *op. cit.* [56] 2000, y en Bußman. *Lexikon der Sprachwissenschaft* [28]. Kröner. Stuttgart. 1990, *s.v.* ALLOMORPH y MORPH.

[3]Por ejemplo, la forma ~a será una marca de género en sustantivos o una de persona y modo (indicativo o subjuntivo) en formas verbales.

'variantes formales' o *morfos*[4], representan a mi juicio —a pesar de su posible polisemia— una abstracción más apta de ser investigada por una máquina que el morfema (por lo menos en un primer acercamiento, como en esta investigación) no sólo porque el morfo es una unidad del habla directamente observable en un corpus, sino también porque concierne a un grupo de formas con uno o varios contenidos cada una y no a un contenido con varias formas. Es decir, mientras que un morfema puede manifestarse en el habla mediante uno o varias formas, cada una constituye un morfo único apto de identificarse y procesarse automáticamente sin tanta supervisión por parte del investigador. De todas maneras, la tarea de decir qué morfos pertenecen a qué morfemas será todavía responsabilidad del lingüista.

El problema principal con una investigación morfológica automática es que las unidades mínimas de la morfología son unidades de significado y los métodos automáticos todavía no son muy aptos para la investigación del significado. De hecho, los criterios semánticos para determinar morfemas son subjetivos y varían de lingüista a lingüista.

De entre todos los fenómenos que se han concebido en la morfología[5] el de afijación ha sido uno muy popular. Se refiere a los procesos de formación de palabras[6] mediante la aplicación

---

[4]Véase Charles Hockett, "Linguistic Elements and their Relations" [68], *Language* 37 (1961), pp. 29-53; *Language* 23 (1947), pp. 321-343.

[5]Para diversas descripciones de tipos de morfemas (según se conciban como objetos, reglas, procesos, etc.), véanse Nida, *Morphology* [110], The University of Michigan Press, Ann Arbor, 1967 [1949], pp. 68-77; Sapir, *El Lenguaje* [122], Fondo de Cultura Económica, México, 1992 [1921], pp. 77-94; Spencer, *Morphological Theory. An Introduction to Word Structure in Generative Grammar* [129], Basil Blackwell, Cambridge, 1991, pp. 4-20; Anderson, *op. cit.* [8] 1994, pp. 48-66; Bergenholtz y Mugdan, *Einführung in die Morphologie* [17], Kohlhammer, Stuttgart, 1979, pp. 58-73.

[6]Básicamente procedimientos que las modifican concatenándoles otros signos o reemplazándolas parcialmente. Como en la mayoría de los fenómenos morfológicos, no siempre hay un consenso en el significado de los términos. Por ejemplo, Andrés Bello utilizó el término 'afijo' para designar los pronombres átonos proclíticos del español, pero no los enclíticos: "Cuando preceden se llaman *afijos*; cuando siguen, *enclíticos*, que quiere decir *arrimados*, porque se juntan con la palabra precedente, formando como una sola dicción. Así se dice: *me parece* o *paréceme*; *os agradezco* o *agradézcoos*; *le* o *lo traje*, y *trájele* o *trájelo*;" etc., Bello, *Gramática de la lengua castellana* [16], Editorial Sopena, Buenos Aires, 5ª ed., 1953, §280 (p. 102); también

de ciertos tipos de signos mínimos llamados afijos a otros llamados raíces.

Aunque estos procesos en sí no son el objeto de investigación en este trabajo, las unidades en cuestión sí, el afijo (como lo indica el nombre del capítulo), en parte porque los morfos de tipo afijal se conciben como elementos desgastados fonética y semánticamente y, para motivos de descubrimiento de morfos afijales, este desgaste semántico representa una gran ventaja, ya que implica menos juicios subjetivos y cualitativos sobre el significado de las formas que tenga que llevar a cabo el investigador.

Por razones similares, este trabajo se abstiene de investigar a las raíces que, consiguientemente, deben cargar la mayor parte de los significados. De todas maneras, al contar con mejores técnicas para descubrir afijos, valdría la pena estudiar, en el marco de un estudio de raíces, el conjunto de elementos que quedan una vez que se han eliminado los afijos. Una última observación con respecto al concepto de raíz en este experimento: para hablar de éstas, aquí utilizaremos también el término 'base' ya que este último se refiere a segmentos que pueden estar constituidos por una combinación de raíces y afijos.

Pero la mayor ventaja de los afijos es que, como se ha venido diciendo desde la introducción, éstos se prestan a una formalización lo suficientemente precisa como para descubrirlos automáticamente[7]. Por ejemplo, aquí se caracterizará al afijo (y por exclusión a la base), como un signo que aparece adherido a la derecha (sufijo) o izquierda (prefijo) de otro signo (la base) y que[8]:

---

§905 (p. 286).

[7]Esto no excluye que otros tipos más problemáticos de signos mínimos, como los involucrados en el proceso de composición, también puedan procesarse computacionalmente, al menos parcialmente.

[8]Nótese que, por comodidad, se tratarán solamente sufijos y prefijos. Quedan pendientes el estudio de otros tipos, tales como los circunfijos y los cambios vocálicos de algunas raíces verbales en español.

1. no ocurren aislados (son parte de las palabras).

2. ocurren en contextos similares (en *muchos* vocablos de relativamente baja frecuencia).

3. tienen significados más generales (por ej., gramaticales) que modifican al significado central de la palabra[9],

4. tienden a ser cortos (limitados en cuanto a repertorio fonológico).

Esta caracterización será nuestra definición de trabajo de afijo[10]. En esencia será una nueva premisa que servirá de base para formalizar y cuantificar la unidad *afijo*. La estrategia escogida para esto consiste en aplicar las medidas involucradas en el cálculo cuantitativo de la propiedad abstracta de ser afijo, según la hipótesis de ***afijalidad*** presentada en la introducción. Primero, la noción tradicional de cuadro (ver más adelante en la página 104) nos proporciona un mecanismo para acercarse al punto 1. Segundo, el principio de economía es ideal para el punto 2. Tercero, aunque la entropía no es una medida de significado, como medida de información bien puede aplicarse para el punto 3. En cuanto al punto 4, aunque no deja de ser un criterio útil para el análisis morfológico, se decidió no tomarlo en cuenta en esta experimento porque, primero, la posibilidad de detectar secuencias de afijos adheridas a una base hace complicado el cálculo de una posible medida promedio de longitud de afijos cuando se considera a la secuencia completa como algo de caracter afijal con respecto a la base. Además, el desgaste fonológico no parece ocurrir ni instantánea ni inmediatamente después de que un segmento adquiera carácter afijal (considérese la longitud del sufijo ~mente). Es

[9]Siguiendo la distinción sapireana de carácter difuso entre contenido material y contenido relacional, los afijos tienden al contenido relacional (hay que notar que lo relacional opuesto a lo material puede variar de lengua a lengua), véase Bybee, *Morphology. A Study of the Relation between Meaning and Form* [29]. John Benjamins, Amsterdam, 1985, p. 7.

[10]Nótese que se excluyen fenómenos como (además del problema de los alomorfos que tratamos arriba) la fusión de bases y afijos (véase Bybee, *op. cit.* [29] 1985, p. 4-7; Anderson pone especial énfasis en el problema que este fenómeno implica para el descubrimiento automático de morfemas, *op. cit.* [8] 1994. pp. 389-391).

decir se trata más de una consecuencia a largo plazo y no tanto de una característica medular de esta unidad morfológica.

Una distinción interesante al interior del concepto de afijo (que probablemente también se pueda investigar formalmente) es aquella entre afijos de derivación y de flexión. Por ejemplo, Greenberg plasmó en sus universales el hecho de que la flexión aparece al exterior de la palabra y que los afijos de derivación aparecen entre los de flexión y la raíz. También se puede sacar provecho de la idea de que los de derivación son más léxicos y los de flexión más sintácticos[11]. Tampoco podemos ignorar lo productiva que podría ser una investigación automática y cuantitativa de los paradigmas, sobre todo de flexión (tema, como se vio arriba, inaugurado por los rusos en los años sesenta). Lo importante es notar que las ideas utilizadas en este trabajo son apenas algunos ejemplos de esquemas explorables mediante el análisis automático de córpora.

2.2 Cuestiones preliminares

En esa sección, se presentan dos dispositivos indispensables para el desarrollo de los experimentos del resto del capítulo: uno abstracto pero formal —el aparato notacional— y otro concreto pero intangible (en el sentido literal de la palabra): las rutinas y estructuras de información que se construyeron para llevar a cabo esta investigación.

[11] Bybee arguye que hay una escala continua entre los polos de derivación y flexión, siendo un extremo más léxico y el otro más sintáctico, *op. cit.* [29] 1985, pp. 81-109.

2.2.1 Nociones formales preliminares

Esta subsección presenta el aparato notacional en que se basa la explicación de las ideas de las secciones siguientes, sobre los métodos de segmentación de vocablos. Se trata de formalizar los procedimientos de extracción de información para, por un lado, facilitar su implementación y, por otro, simplificar su exposición y entendimiento.

Sea Ψ una secuencia de ocurrencias de palabras (por ejemplo, un corpus) de tamaño ξ,

$$\Psi = o_1, o_2, o_3, \ldots, o_\xi$$

donde cada palabra puede ser idéntica a cualquier otra excepto en su posición en la secuencia. Sea Φ el conjunto de Ω pares ordenados:

$$\Phi = \{\langle v_1, f_1\rangle, \langle v_2, f_2\rangle, \langle v_3, f_3\rangle, \ldots \langle v_\Omega, f_\Omega\rangle\},$$

donde cada v_i es miembro también del conjunto V de vocablos (o formas de palabras). $V = \{v_1, v_2, v_3, \ldots, v_\Omega\}$, y f_i del conjunto F de valores de sus frecuencias en Ψ. $F = \{f_1, f_2, f_3, \ldots f_\Omega\}$; $V \subset \Phi$ y $F \subset \Phi$. Así, la suma de estas frecuencias es igual al tamaño de la secuencia Ψ:

$$\sum_{i=1}^{\Omega} f_i = \xi$$

Cada segmentación posible de cada vocablo v_i se puede representar mediante '::': esto es, si dividimos un vocablo en dos segmentos, sea '$a::b$' la representación de esa segmentación, donde 'a' corresponde al segmento de la izquierda y 'b' al de la derecha. Para recordar que cada segmento es parte de un vocablo particular, incluiré el índice i que corresponde al vocablo examinado, de tal manera que $a_i::b_i \equiv v_i$.

Además, sea j otro índice que indique cada segmentación del vocablo v_i (es decir, la columna donde se dividen los segmentos de ese vocablo). Así, si v_i tiene m_i caracteres de longitud, entonces contiene $m_i - 1$ segmentaciones posibles, $a_{i,j}::b_{i,j}$, donde $j = \{1, 2, ... m_i - 1\}$. Por ejemplo, si tenemos el vocablo v_x, sus segmentaciones se representarán como en la figura 2.1.

Figura 2.1: Representación de las segmentaciones posibles de un vocablo x (v_x).

$a_{x,1}::b_{x,1}$ (e::jemplo)

$a_{x,2}::b_{x,2}$ (ej::emplo)

$a_{x,3}::b_{x,3}$ (eje::mplo)

$\vdots$

$a_{x,m_x-1}::b_{x,m_x-1}$ (ejempl::o)

Definamos el proceso de *alternancia* como aquel en el que, dado un vocablo v_i dividido en dos segmentos, se buscan todos los vocablos en V que comparten con v_i uno de esos segmentos, el que no alterna. Es decir, si consideramos *fijo* al segmento $a_{i,j}$ de v_i, hay un conjunto de segmentos, al cual pertenece por lo menos $b_{i,j}$, que se combinan con el primero para formar vocablos del corpus. De manera similar, si *fijamos* el segmento $b_{i,j}$, hay un conjunto de segmentos (al que pertenece $a_{i,j}$) que alternan a la izquierda. Así, aunque no se diga explícitamente en adelante, decir que uno de los extremos del vocablo alterna, significa que el otro permanece estático o fijo[12].

[12]De Kock no define explícitamente el término *alternar* pero hace uso de él con un significado similar al que aquí se define (*op. cit.* [82] 1978, p. 17). Sin embargo, el concepto de 'alternación' (*alternation*) tiene en morfología varias acepciones, véase Matthews, *Morphology* [95]. Cambridge University Press. Cambridge, 1991, pp. 114-119. También Wells, se ocupa de esta palabra en el marco de la morfofonémica ("Automatic Alternation" [135], *Language* 25 (1949), pp. 99-116), pero se pueden utilizar sus términos *communis* (la parte que comparten un conjunto de formas) y *propria* (la parte propia a cada forma, que no comparte con las

Sea $A_{i,j}$ el conjunto de segmentos que alternan a la izquierda del segmento $b_{i,j}$, es decir, el conjunto de segmentos encontrados al dejar alternar $a_{i,j}$ ($a_{i,j} \in A_{i,j}$); y sea $B_{i,j}$ el conjunto que alterna a la derecha de $a_{i,j}$ (por lo que $b_{i,j} \in B_{i,j}$). Entonces, $|A_{i,j}|$ será el número de miembros de $A_{i,j}$ y $|B_{i,j}|$ el de $B_{i,j}$.

2.2.2 Rutinas y estructuras de datos

En este espacio se describen brevemente las rutinas y estructuras de información que se construyeron para llevar a cabo la investigación del presente capítulo. Esto es necesario para entender cómo y dónde se almacenan los índices que se verán en la próxima sección. Al igual que en la presentación del aparato notacional de la subsección precedente, la idea es formalizar los procedimientos de extracción de información para, por un lado, facilitar su implementación y, por otro, simplificar su exposición y entendimiento.

El primer conjunto de rutinas que se elaboró corresponde al preprocesamiento del corpus. En el primer apéndice, "El *Corpus del Español Mexicano Contemporáneo*", hay una sección que describe cómo se preparó el corpus para eliminar las ambigüedades entre caracteres y hacerlo más manejable al quitarle información no directamente relacionada con esta investigación, tal como las claves de cada línea, las marcas gramaticales asociadas a cada palabra gráfica y los separadores entre estos datos.

El siguiente paso fue construir un programa que genera, a partir del corpus, la cadena de palabras gráficas en orden secuencial. Se trata de un segmentador de texto que se conoce

---

demás) para describir lo que hemos de entender por alternancia en este trabajo: los segmentos alternantes son el conjunto de partes propias (únicas) que están en uno de los extremos de un conjunto de vocablos, mientras que el segmento fijo es el otro extremo, que es la parte *communis* de todos ellos, p. 104.

en inglés como *tokenizer* y que aquí llamaremos fichador. Hacer operativo un programa de este tipo no es una cuestión tan sencilla como parecería a primera vista, sobre todo porque nuestra premisa inicial de definir a la palabra como una secuencia de caracteres entre dos espacios, si bien sólo es operativa, no es lo suficientemente precisa[13].

Por ejemplo, aunque en los términos estrictos de nuestra definición los números también son palabras, no está claro que eso convenga a nuestros propósitos (especialmente cuando lo que se quiere es descubrir afijos[14]). Otro problema es la naturaleza de ciertos caracteres no alfanuméricos. Por ejemplo, los guiones pueden ocurrir tanto al interior como al exterior de las palabras gráficas. Entre éstas pueden marcar una pausa o tener una función parentética, pero algunos dividen palabras porque empiezan al final de un renglón y terminan al principio del siguiente[15] ('diver-' al de una línea seguida de otra que empieza con 'gencia'). Además, los guiones también separan palabras gráficas distintas ('Orizaba-Cempoaltépetl', 'Nonoalco-Tlatelolco', 'flor-de-mayo', 'los-ruidos', 'no-hijos' y composiciones como 'caballeros-águila', 'perro-lobo', 'tren-hospital', etc.) y, lo que es peor, muy a menudo ocurren entre segmentos de una misma palabra ('semi-ortopédica', 'ex-niño', 'cha-cha-cha', 'neo-ovidiana', etc., etc.)[16].

Luego está el problema de la ambigüedad de los puntos: a veces no sólo separan ora-

---

[13]Los problemas que enfrenta un programa de este tipo con respecto a la definición de palabra en inglés son bien conocidos y se reseñan, por ejemplo, en Mason, *Programming for Corpus Linguistics* [94]. Edinburgh University Press, Edinburgh, 2000, pp. 134-135, y Manning y Schütze, *op. cit.* [93] 1999, pp. 125-127.

[14]De esta manera, en este capítulo los números no fueron considerados palabras. No así en la investigación del siguiente capítulo, donde se observó que los números pueden funcionar como adjetivos cuantificadores de sustantivos (`'...tenía 12 años'`) o, incluso, sustantivos (`'Pablo bajaba de la 38'`).

[15]Las ocurrencias de estas palabras se buscaron automáticamente y como eran tan pocas simplemente se eliminaron.

[16]Por esto último, las ocurrencias de guiones rodeados de letras se consideraron separadores de segmentos de palabras, lo que inevitablemente resultó en palabras tan largas como `'cobra-vida-propia-y-se-independiza-del-escritor'` (en ©1290130090).

ciones, sino que marcan abreviaturas o, incluso, tienen ambas funciones[17]. Similarmente, los apóstrofos también son un problema porque a veces no son los símbolos ortográficos que indican la elisión de letras, sino comillas sencillas que cierran un segmento del texto[18].

También la cuestión de convertir los símbolos ortográficos a caracteres representativos de la fonología del español es un asunto que incide en la construcción de un fichador. Como se verá más adelante, se especificaron cuatro filtros para diferentes tipos de correspondencias[19]. No está de más decir que muchas veces la aplicación de estos filtros resultó en formas homógrafas que incidían en el tamaño de los vocabularios obtenidos del corpus (los segmentos 'ha' y 'a' se convirtieron simplemente en 'a'; 'parece' y 'párese' en 'parese'; etc.).

Aunque por todas estas complicaciones el carácter de palabra gráfica de las *fichas* generadas por el fichador es a veces peculiar, la cadena de segmentos que se obtiene sirve de base para empezar con la investigación.

Así, una vez que se tiene un esquema de determinación de palabras de un corpus, se procede a construir una estructura que las contenga. En las primeras versiones de los programas desarrollados para procesar la nomenclatura del DEM[20], las palabras simplemente servían para construir listas de vocablos con información cuantitativa asociada a cada letra de cada vocablo. Cada experimento tenía asociadas dos listas: una ordenada alfabéticamente

---

[17]En el apéndice se describe el procedimiento de descubrimiento de abreviaturas que se aplicó al *CEMC*, con miras a desambiguar los tipos de puntos con que se encuentra el fichador.

[18]El fichador lleva una contabilidad de todo lo que funge y puede fungir como marca de paréntesis o de inicio y final de algún tipo de secuencia, esto es, '"', ')', '(', '¿', etc. y '-' (guión), ''' (apóstrofo). De esta manera, se tiene información sobre la posibilidad de que un guión o un apóstrofo sea parte de una palabra o cierre algún tipo de secuencia.

[19]La tabla A.7 del apéndice muestra las reglas que se aplicaron en tres de los cuatro filtros (en el cuarto se conservaron los caracteres ortográficos).

[20]Véase Medina art. cit. [96] 1996.

de izquierda a derecha y otra de derecha izquierda (como en los diccionarios inversos) para calcular para cada par de letras de cada vocablo en cada dirección los índices cuantitativos que se describirán en la próxima sección.

En las versiones posteriores, al tomar en cuenta la totalidad de las palabras del corpus, se hizo necesario recurrir a estructuras arbóreas para ahorrar espacio y acelerar los procedimientos. Entonces, cada vez que se echaba a andar el procedimiento, en lugar de listas se construían dos estructuras arbóreas cuyos nodos corresponden a los caracteres de los vocablos. De la raíz del primero de los árboles se llega a los nodos de la primera columna de todos los vocablos (y de estos a los de la segunda columna y así hasta el final) y de la raíz del segundo se llega a los nodos de sus últimas letras (que a su vez dan acceso a los de las penúltimas y éstos a los de las antepenúltimas y así hasta el principio)[21].

Para construir las estructuras arbóreas se escribieron varias rutinas para iniciarlas, almacenarlas en disco duro, insertarles palabras (agregar nodos), eliminarlas (quitar nodos), etc. que mantenían al corriente la información cuantitativa (frecuencias de segmentos en palabras y vocablos —*tokens and types*). Además del espacio para almacenar las frecuencias, que se actualizan automáticamente con cada nueva palabra que se inserta, cada nodo tiene campos para cada uno de los índices cuantitativos que se describen en la próxima sección. Gran parte del cálculo de éstos se lleva a cabo en una etapa posterior a la construcción de los árboles.

La etapa posterior a la construcción de dichas estructuras se lleva a cabo en lo que podemos llamar un segmentador que simplemente propone fronteras morfológicas donde ocurren los

[21]Como se verá en el próximo capítulo, se trata de cadenas de Markov de varios órdenes. La figura 3.4 contiene una representación de esto (página 184). La definición formal de este tipo de estructura se presenta a partir de la página 192.

valores más altos de los índices que veremos. Mediante una interfaz sencilla, el usuario puede designar las palabras que se analizarán. El programa toma cada palabra y, después de calcular los valores pertinentes ofrece una secuencia de caracteres que corresponden al vocablo y sus segmentaciones (letras y símbolos representativos de los valores más altos de los índices calculados[22]). Por último, como veremos después, también se construyeron rutinas para compilar en catálogos (otras estructuras arbóreas) los segmentos con los valores más altos. Entre estas rutinas hay procedimientos para agregar, eliminar, ordenar y mostrar listas de los miembros de esos catálogos. La formalización de estos catálogos aparece más adelante en este capítulo (a partir de la página 123).

2.3 Índices para cuantificar la *afijalidad* de una segmentación

En este apartado se examinan varias técnicas para segmentar palabras. La idea es determinar las más apropiadas en la separación de raíces y afijos con el objeto de construir un catálogo de estos últimos. En esencia, se trata de los métodos de descubrimiento morfológico que se presentaron arriba y que no requieren conocimiento previo de los límites entre morfemas, esto es, estadísticas de digramas, cuentas de fonemas anteriores y posteriores, principio de economía y número de cuadros. Para explorar estas diferentes estrategias de segmentación morfológica, se calcularon diferentes índices para cada segmentación de cada vocablo del conjunto V, con el objeto de compararlos y construir con ellos una herramienta que sirva para determinar si una segmentación dada corresponde o no a la frontera entre dos morfos.

[22]En la tabla B.3 del apéndice (página 377) se especifican los símbolos que corresponden a cada índice.

2.3.1 Número de cuadros

Este apartado se ocupa del concepto de *cuadro* de la lingüística estructural como medida de la validez de una segmentación morfológica. Greenberg lo caracteriza como un conjunto de palabras que

> exists when there are four expressions in a language which take the form AC, BC, AD, BD. An example is English *eating:walking::eats:walks*, where A is *eat-*, B is *walk-*, C is *-ing*, and D is *-s*. One of the four members may be zero, as in *king:kingdom::duke::dukedom*, where C is zero.[23]

Así, un cuadro será aquí un conjunto de cuatro segmentos de vocablos, dos de la izquierda (a_1 y a_2) y dos de la derecha (b_1 y b_2) que combinados (los de la izquierda con los de la derecha, esto es, $a_1::b_1$, $a_1::b_2$, $a_2::b_1$, $a_2::b_2$) resulten en vocablos presentes en el conjunto V. Uno de los segmentos podrá ser la cadena nula de caracteres, ø, para permitir cuadros tales como $\{in::cauto.\ in::feliz,\ ø::cauto,\ ø::feliz\}$. Esta estructura combinatoria puede variar de acuerdo al número de elementos que se requieran. Por ejemplo, pueden requerirse seis, de tal manera que en lugar de cuadro se tiene un *hexágono* (seis segmentos contenidos en seis palabras), o para córpora muy pequeños, se puede relajar el requisito para aceptar cuadros incompletos[24]. En la tabla 2.1 se ilustran estas combinaciones:

De esta manera, se pueden contar los cuadros posibles de cada segmentación examinada al permitir que *alternen*, uno a la vez, los segmentos a cada lado de dicha segmentación. Es decir, cada segmentación posible de cada vocablo v_i se examina para contar el número de

[23] Joseph H. Greenberg, *op. cit.* [58] 1957, p. 20.

[24] "Un inventario restringido pide una regla elástica, un inventario ampliado admite una regla severa. La diferencia radica en que en un vocabulario limitado todas las posibilidades de realización, efectivas en el conjunto de la lengua, y que precisamente motivan la segmentación morfológica y la autorizan, no se hallan siempre representadas", de Kock y Bossaert, art. cit. [81] 1974 p. 195.

Tabla 2.1: Tipos de combinaciones de segmentos

cuadro	cuadro incompleto	hexágono
A::a	A::a	A::a, A::b, A::c
A::b	A::b	B::a, B::b, B::c
B::a	B::a	C::a, C::b, C::c
B::b	B::c	

cuadros documentables allí, al buscar las posibles combinaciones en el conjunto V. Llamemos a este número $c_{i,j}$ (el número de cuadros encontrados en el segmento j del vocablo i).

Un problema importante es que no cualquier cuadro posible es aceptable. Para evitar cuadros tales como {*k*::*apasidád*, *r̄*::*apasidád*, *k*::*apás*, *r̄*::*apás*}, pero permitir aquellos que verdaderamente atestiguan una frontera morfológica (según la evidencia del corpus), es necesario aplicar alguna prueba de correspondencia de significado[25]. La naturaleza de este experimento, sin embargo, impide que se aplique alguna prueba de este tipo. De todas maneras, con base en nuestra premisa que establece que los errores son inevitables, pero detectables gracias a sus bajas frecuencias en comparación con los fenómenos sistemáticos del lenguaje, me atreveré a asumir que serán muy pocos los cuadros correspondientes a una segmentación no morfológica en comparación con la cantidad de cuadros que atestigüen una sí morfológica.

Por último, en este experimento se aplicó una restricción importante en el conteo de cuadros que se pudieran detectar para cada segmentación. Ya que se trata de descubrir afijos, se contaron solamente aquellos cuadros donde los segmentos alternantes guardaban una relación de pocos y muy frecuentes con respecto al segmento fijo o, viceversa, cuando el segmento fijo pertenecía a un conjunto más pequeño de segmentos más numerosos que el conjunto de los segmentos alternantes.

[25]Greenberg, *op. cit.* [58] 1957, p. 23.

2.3.2 Entropía

Las intuiciones detrás del método de Harris —en el sentido de tomar el número de signos que preceden o siguen una segmentación para descubrir una frontera morfológica— tienen mucho sentido al considerar el fenómeno de afijación (más que cualquier otro de los fenómenos morfológicos). Considerémoslo en términos de la teoría de la información. Es decir. en términos de entropía.

Greenberg presenta esta idea con mucha claridad: "both in the technical sense of information theory and in the nontechnical meaning of information, the utterance of a member of a root class of morphemes gives more information"[26]. De esta manera, si se supone que un afijo contiene información predominantemente gramatical —en contraste con las bases. que contienen mucho más (la ocurrencia de una base en particular nos debe sorprender mucho más que la de un afijo)—, entonces podemos esperar que un pico de entropía al interior de una palabra señale el principio de una base. mientras que el lugar donde ocurra la entropía más baja será el inicio de un sufijo.

Tómese. por ejemplo, el conjunto Φ de vocablos y sus frecuencias. Existen cuando menos dos maneras de calcular la entropía de cada segmentación de cada vocablo, dependiendo de qué frecuencias son las que se toman en cuenta. Es decir, los vocablos tienen una frecuencia en el corpus, pero los segmentos de ellos se repiten en otros vocablos, por lo que esos segmentos tienen una frecuencia de aparición en los vocablos y otra en el corpus. En este experimento se tomaron en cuenta solamente las frecuencias en los vocablos.

---

[26]Greenberg, *op. cit.* [58] 1957, p. 91.

Considérese que al examinar una segmentación y determinar el conjunto de segmentos que alternan en alguno de los extremos del vocablo en cuestión y que cada uno de esos segmentos alternantes tiene una frecuencia dada. se pueden calcular probabilidades para cada uno de estos segmentos; es decir, si imaginamos un depósito hipotético de segmentos que alternan. podemos calcular las probabilidades que tiene cada uno de ser escogido al azar. Así, a partir de un vocablo $a_{i,j}::b_{i,j}$, podemos imaginarnos al conjunto $B_{i,j}$ como un depósito de segmentos con posibilidades de ser seleccionados. La probabilidad de cada segmento en ese conjunto se dará de la siguiente manera:

$$0 \leq p(b_{k,j} \mid a_{i,j}) = \frac{f(b_{k,j})}{f(a_{i,j})} \leq 1, \qquad b_{k,j} \in B_{i,j}, \quad k = 1, 2, 3, ...|B_{i,j}|$$

donde la suma de estas probabilidades será 1:

$$\sum_{k=1}^{|B_{i,j}|} p(b_{k,j} \mid a_{i,j}) = 1$$

Por ejemplo, tomemos el vocablo 'previamente' y examinemos la primera segmentación posible: *p::rebiamente*. En nuestra lista de vocablos, hay 7206 que empiezan con *p*−. es decir. $f(a_{i,j}) = 7206$, que es también el número de elementos del conjunto $B_{i,1}$ (v_i = 'prebiamente'). La tabla 2.2 muestra el número de formas en $B_{i,1}$ que empiezan con una letra en común. Nótese que la mayoría de formas que aparecen después de *p*− empiezan con vocal o consonante laminal[27]. Todas las demás (pb∼..., pc∼..., pd∼..., etc.) corresponden a siglas, abreviaciones. palabras extranjeras, símbolos matemáticos o químicos, etc. El total de la última columna nos muestra la entropía en esa segmentación.

Entonces, basándose en la fórmula presentada arriba (página 84), se puede calcular la

[27] Algunas formas con 's' (ps∼icología, ps∼íquico, etc.).

Tabla 2.2: Entropía de la segmentación $p::B_{i,1}$

$a_{i,j}::B_{i,j}$	formas	$p(b_{k,j} \mid a_{i,j})$	-p×log(p)
p::a	1365	0.189425	0.31518
p::b	1	0.000138773	0.00123268
p::c	2	0.000277546	0.00227297
p::d	1	0.000138773	0.00123262
p::e	1396	0.193727	0.317965
p::f	1	0.000138773	0.00123268
p::g	1	0.000138773	0.00123268
p::h	3	0.00041632	0.00324066
p::i	511	0.0709131	0.187657
p::j	1	0.000138773	0.00123268
p::k	1	0.000138773	0.00123268
p::l	384	0.0532889	0.156245
p::m	2	0.000277546	0.00227297
p::n	2	0.000277546	0.00227297
p::o	835	0.115876	0.24974
p::p	3	0.00041632	0.00324066
p::r	2184	0.303081	0.361804
p::s	73	0.0101304	0.0465211
p::t	7	0.000971413	0.00673846
p::u	407	0.0564807	0.162317
p::v	2	0.000277546	0.00227297
p::x	1	0.000138773	0.00123268
p::y	1	0.000138773	0.00123268
p::z	2	0.000277546	0.00227297
p::-	13	0.00180405	0.0113975
p::'	6	0.000832639	0.00590417
p::ø	1	0.000277546	0.00227297
total	7206	1.0	1.85039 bits

entropía de la segmentación j dada la parte izquierda del vocablo i:

$$H(i,j)^{izq} = -\sum_{x=1}^{f(a_{i,j})} p(b_{x,j} \mid a_{i,j}) \times log_2(p(b_{x,j} \mid a_{i,j}))$$

En esencia, esta entropía mide la información, en el sentido técnico de la palabra, que la segunda parte del vocablo debe proporcionar para reducir la incertidumbre generada por las alternativas posibles. Como ya se ha dicho, si definimos a un afijo como una unidad con relativamente poca información, se puede intuir que una medida baja, en relación a las otras segmentaciones de la palabra, indica el inicio de un sufijo, mientras que la medida más alta indica el inicio de una base.

Por otra parte y aunque sea poco intuitivo, al examinar la entropía de la misma segmentación pero en sentido contrario (al fijar $b_{i,j}$ para obtener $A_{i,j}$ y medir las probabilidades de sus miembros), los datos muestran que la situación es similar, pero, puesto que el flujo de información va de izquierda a derecha, los valores altos corresponden a fronteras de sufijos y los bajos a las de prefijos. Esto no es ni tan extraño ni tan nuevo, por ejemplo, Harris —aunque no habla de entropía— encuentra que las cuentas de fonemas anteriores a una segmentación son tan buenos indicios de frontera morfológica como las de los posteriores. En cuanto a la entropía concretamente, también Hafer y Weiss, art.cit. [60] 1974 notaron que *en reversa* es tan buen indicador morfológico como de izquierda a derecha.

Tabla 2.3: Valores de entropía en cada segmentación del vocablo 'aparecer'.

	A	P	A	R	E	S	E	R
izq.-der.		**2.792**	1.818	1.63	1.298	1.27	0.9497	1.303
der.-izq.		1.277	0.8018	1.619	2.125	1.56	**2.516**	1.193

De hecho, puesto que los picos de entropía señalan el principio de bases después de un prefijo, los picos de entropía *al revés* marcan el inicio de un sufijo. Es más, en este experimento, las entropías *en reversa* resultaron mejores indicadores de sufijos que los valores más bajos de la entropía de izquierda a derecha. Similarmente, los valores máximos de ésta última resultaron ser mejores indicadores de fronteras entre prefijos y bases. Por ejemplo, en la tabla 2.3 aparecen los valores de entropía calculados en cada segmentación del vocablo 'aparecer' ([*apareser*]) en ambas direcciones. Nótese que la entropía del prefijo a~ no corresponde al valor más pequeño de la serie que va de derecha a izquierda, sino al mayor de la de izquierda a derecha.

Una posible explicación es que los afijos, no por cargar relativamente menos información,

dejan de cargarla: un signo sin contenido es una contradicción de términos. De hecho, puede haber entropías más bajas al interior de ellos. Así, si la entropía más baja de izquierda a derecha no marca el inicio de un sufijo (aunque la más alta sí marque el fin del prefijo), entonces nada más estamos en posición de decir que el sufijo no empieza donde está la entropía más baja. Por otra parte, saber dónde empieza o termina una base es un muy buen indicador para suponer el final o principio de un afijo. Lo importante es recalcar que el afijo no es la unidad con menos información, sino que la base es la entidad con el más alto contenido de ésta.

Principalmente por esta razón, hay que tener cuidado en lo que se refiere a combinar los valores de una dirección con los de la otra, por ejemplo, substrayendo uno del otro. Si por un lado, Harris propuso el uso de cuentas de fonemas anteriores principalmente para cotejar los resultados derivados de las cuentas de los fonemas posteriores, por el otro, Hafer y Weiss combinaron estas cuentas de diferentes maneras sin obtener resultados alentadores. De hecho, los resultados fueron más bien mediocres, por ejemplo, al sumarlos. Además, señalaron que "the union of the measures produces too many incorrect cuts, while the intersection of the methods is too restrictive"[28]. De hecho, como se verá más adelante, los valores de entropía de cualquiera de las dos direcciones son mejores pistas por separado que, por ejemplo, mediante la substracción del valor de una dirección menos el de la contraria.

2.3.3 Índice de economía

En este apartado se elaboran con detalle las ideas que, basados en el principio de economía

[28] Hafer y Weiss, art. cit. [60] 1974, p. 378.

(véase la página 87), Josse de Kock y Walter Bossaert aplicaron en su método de segmentación morfológica.

En los años setenta, de Kock y Bossaert desarrollaron un método para determinar empíricamente fronteras morfológicas entre bases y afijos. El procedimiento está basado, como se dijo anteriormente, en el principio de economía: el número de signos en todos los niveles del lenguaje tiende a ser menor que el número de cosas nombradas: de tal manera que cada signo puede utilizarse en diferentes contextos sin crear ambigüedad. Así, mientras menos signos de más frecuencia existan en el nivel morfológico, que den lugar a más signos (de baja frecuencia) del nivel sintáctico, la lengua será más económica. Este mecanismo sencillo nos proporciona un criterio para determinar los signos morfológicos.

De esta manera, si se divide un vocablo v_i en dos segmentos, $a_i::b_i$, y uno de éstos ocurre en otros muchos vocablos, mientras que el otro ocurre en unos cuantos y, si el primero pertenece a un conjunto pequeño de segmentos muy frecuentes, mientras el segundo pertenece a un conjunto potencialmente infinito de segmentos de baja frecuencia, se puede proponer un corte morfológico entre esos dos segmentos. Es más, el primero tendrá que ser un afijo y el segundo una base.

Por ejemplo, tómense los segmentos de la figura 2.2. Cada segmento de la izquierda se combina con cada segmento de la derecha para formar vocablos españoles (*compra, comprada, comprado, comprando, ... compró; ... canta, cantada, ... cantó; ... controló;* .... etc.). Nótese que las formas de la derecha constituyen un conjunto B más bien pequeño de formas muy frecuentes y las de la izquierda uno A de muchísimos más miembros (un número potencialmente infinito) que son relativamente menos frecuentes (aparecen en menos vocablos

Figura 2.2: Combinaciones de segmentos de la izquierda y de la derecha

A	B
compr	a
cant	ada
alivi	ado
rest	ando
ray	ar
sum	aron
seleccion	aste
⋮	⋮
arrest	es
elabor	é
nad	o
anhel	ó
contrat	
am	
apel	
mand	
colabor	
control	
⋮	
∞	

que las primeras). Esto es una pista muy razonable de que los segmentos del conjunto B son sufijos. Así, al comparar los tamaños de estos conjuntos, se puede argüir que la segmentación es morfológica.

Recuérdese que $A_{i,j}$ es el conjunto de segmentos que alternan a la izquierda del segmento $b_{i,j}$, es decir, el conjunto de segmentos encontrados al dejar alternar $a_{i,j}$ ($a_{i,j} \in A_{i,j}$); y que, similarmente, $B_{i,j}$ es el conjunto que alterna a la derecha de $a_{i,j}$ (por lo que $b_{i,j} \in B_{i,j}$). Entonces, $|A_{i,j}|$ es el tamaño de $A_{i,j}$ y $|B_{i,j}|$ el de $B_{i,j}$. De Kock y Bossaert llamaron a estos números m_g y m_d[29].

Para afinar sus resultados, de Kock y Bossaert aplican varias restricciones a los miembros

[29]Uno para la izquierda (*gauche*) y otro para la derecha (*droit*); véase *op. cit.* [82] 1978, p. 18.

de estos conjuntos. Se trata de eliminar de ambos lados cualquier segmento apto de ser una base. La idea es no contar los segmentos de menor frecuencia que la de su segmento acompañante (esto es, las bases deben ser mucho menos frecuentes que los afijos que las acompañan). Sea $A^p_{i,j}$ el conjunto de segmentos que hipotéticamente funcionan como prefijos y $B^s_{i,j}$ el de sufijos hipotéticos. Entonces, $A^p_{i,j}$ es miembro de $A_{i,j}$ ($A^p_{i,j} \in A_{i,j}$) y contiene aquellos miembros de $A_{i,j}$ con frecuencias que comparadas con las de los miembros de $B_{i,j}$ delatan una conducta de prefijos. Similarmente, los miembros de $B^s_{i,j}$ son aquellos de $B_{i,j}$ ($B^s_{i,j} \in B_{i,j}$) que funcionan como sufijos. Además, sea $|A^p_{i,j}|$ el número de miembros de $A^p_{i,j}$ y $|B^s_{i,j}|$ el número de miembros de $B^s_{i,j}$.

Pero antes de eliminar los segmentos aptos de ser bases, hay otras restricciones que aplican de Kock y Bossaert. Consideremos una de las segmentaciones del vocablo específico $a_{i,j}::b_{i,j}$. Para cada uno de sus segmentos, hay un conjunto de formas alternantes en el extremo opuesto: $a_{i,j}::B_{i,j}$ y $A_{i,j}::b_{i,j}$. Con frecuencia se dará la situación en que $B_{i,j}$ contenga varios segmentos que empiezan con una letra en común o que $A_{i,j}$ contenga varios elementos que comparten la misma letra final. Esto significa que esas letras en común son, con toda probabilidad, parte del segmento opuesto (el fijo), puesto que son adyacentes a éste último. En otras palabras, si dentro del conjunto de afijos hipotéticos, hay alguno que aparece con varias bases que comparten el fonema adyacente a éste, se puede suponer que ese fonema es parte del supuesto afijo y no de las supuestas bases. De manera similar, si dentro del conjunto de bases hipotéticas, hay una cuyos afijos correspondientes comparten fonemas adyacentes a ésta, se puede sospechar que esos fonemas pertenecen a la base[30]. Así, los casos de fonemas

[30]De Kock y Bossaert, *op. cit.* [82] 1978, p. 21.

adyacentes a la segmentación, que se repiten en los segmentos que alternan, deben eliminarse de la cuenta. Por ejemplo, en la figura 2.2, las formas *-a*, *-ada*, *-ado*, *-ando*, *-ar*, *-aron*, *-aste* contarían como una sola. Igualmente, *compr-*, *elabor-* y *colabor-* representarían una sola forma[31].

Una última restricción importante[32] consiste en requerir que ambos conjuntos contengan miembros que aparezcan en más de un vocablo (es decir, cada segmento debe ocurrir en por lo menos dos vocablos diferentes). Esto se garantiza al requerir la existencia de por lo menos un cuadro. Pero, debido a la dificultad de determinar un número mínimo de cuadros[33], en experimento llevado a cabo con el *CEMC* se requirió la presencia de sólo uno (la aus de cuadros es un fuerte indicio de que no hay corte morfológico).

La economía de una segmentación se puede cuantificar al comparar los tamaños de los conjuntos modificados con estas restricciones: mientras más grande sea la diferencia en número de segmentos considerados como bases con respecto al número de aquellos asumidos afijos, más económica será la segmentación. Si alternan a la izquierda más segmentos de tipo base ($|A_{i,j}| - |A^p_{i,j}|$) que segmentos de tipo afijo a la derecha ($|B^s_{i,j}|$), tiene sentido considerar sufijo al segmento de la derecha $b_{i,j}$; y al revés, si más segmentos de tipo base alternan a la derecha ($|B_{i,j}| - |B^s_{i,j}|$) que segmentos de tipo afijo a la izquierda ($|A^p_{i,j}|$), tiene sentido suponer que

---

[31] En la notación de de Kock y Bossaert, los tamaños de estos conjuntos modificados corresponden a las medidas M_g y M_d.

[32] De Kock y Bossaert, (*op. cit.* [82] 1978 p. 23) aplicaron otras restricciones que resultaron tener poco impacto. Por ejemplo: requerir que las bases contengan por lo menos una vocal; eliminar afijos conocidos de entre los segmentos examinados como bases; etc.

[33] No es un problema trivial el determinar un límite mínimo de cuadros que atestigüen fronteras morfológicas (en parte porque, como se apuntó arriba, no hay una prueba automática de significado y con frecuencia se observan en segmentaciones no morfológicas cuadros semánticamente inaceptables). Por otra parte, este requisito no es raro, además del de Kock y Bossaert, el procedimiento de Andreev para detectar paradigmas también requiere la presencia de cuadros, *op. cit.* [46] 1996, p.8.).

el segmento de la izquierda $a_{i,j}$ es un prefijo. De esta manera, tenemos esencialmente dos medidas de economía asociadas a una segmentación, dependiendo del tipo de afijo que se hipotetice:

$$k^p_{i,j} = \frac{|B_{i,j}| - |B^s_{i,j}|}{|A^p_{i,j}|} \tag{2.1}$$

donde $k^p_{i,j}$ medirá la economía de la segmentación j del vocablo v_i y tendrá un valor mayor a la unidad cuando su primer segmento, $a_{i,j}$, sea un prefijo o será una fracción cuando el segmento de la derecha $b_{i,j}$ sea un sufijo y, similarmente,

$$k^s_{i,j} = \frac{|A_{i,j}| - |A^p_{i,j}|}{|B^s_{i,j}|} \tag{2.2}$$

donde $k^s_{i,j}$ medirá la economía de la segmentación j del vocablo v_i y tendrá un valor mayor a la unidad cuando $b_{i,j}$ sea un sufijo o será menor a ésta cuando $a_{i,j}$ sea un prefijo. Aunque la notación de de Kock y Bossaert es muy diferente, estos cocientes son básicamente lo que exploraron en los años setenta, así que en adelante nos referiremos alternativamente a ellos como cociente de de Kock-Bossaert o índice de economía.

En el experimento aplicado al *CEMC* se aplicaron a cada segmentación de cada vocablo versiones normalizadas de estos índices (para obtener valores entre cero y la unidad, [0.1]):

$$k^p_{i,j} = 1 - \frac{|A_{i,j}| - |A^p_{i,j}|}{|B^s_{i,j}|} \tag{2.3}$$

para el segmento de la izquierda como prefijo, y

$$k^s_{i,j} = 1 - \frac{|B_{i,j}| - |B^s_{i,j}|}{|A^p_{i,j}|} \tag{2.4}$$

para el de la derecha como sufijo.

2.3.4 Medidas estadísticas

En esta sección se expone la aplicación de las estadísticas de co-ocurrencia de digramas a la lista de vocablos extraída del corpus. Como se apuntó arriba la estadística de digramas nos proporciona varias medidas de no asociación (o independencia) que se han estudiado en el marco de la segmentación morfológica[34].

Para cada segmentación posible de cada vocablo, se puede construir una tabla de contingencia como la tabla 1.7 de la página 73. Por ejemplo, supongamos que queremos examinar la segmentación entre los segmentos previa~ y ~mente del vocablo 'previamente' que cuenta con 58 ocurrencias en el corpus, entonces la tabla de contingencia se verá como aquella en 2.4. La primera columna contiene las frecuencias del segmento ~mente. En el primer renglón las

Tabla 2.4: Tabla de contingencia para el digrama 'previa::mente'.

	::mente	$\overline{::mente}$	**total**
prebia::	f(*prebia*::*mente*)=58	f(*prebia*::$\overline{mente}$)=36	f(*prebia*::)=94
$\overline{prebia}$::	f($\overline{prebia}$::*mente*)=12,190	f($\overline{prebia}$::$\overline{mente}$)=1,939,963	f($\overline{prebia}$::)=1,952,153
total	f(::*mente*)=12,248	f(::$\overline{mente}$)=1,939,999	$T = \sum f(w_i)$=1,952,247

veces que aparece junto a previa~. En el segundo las veces que aparece en otro contexto, es decir, cuando lo que le precede es cualquier cosa otra que previa~: $\overline{\text{previa}\sim}$ (la raya encima significa 'no' o 'ausencia de'). Similarmente, la segunda columna contiene las frecuencias de todo lo que no fue ~mente. En el primer renglón las ocurrencias de previa~ sin ~mente y en segundo la frecuencia de todas las palabras del corpus que no fueron 'previamente'. Luego, con las fórmulas de la tabla 1.8 (página 74) podemos calcular por ejemplo la medida de la

[34]Véase la tabla 1.8 de la página 74.

prueba χ^2:

$$X^2 = \frac{T(f(prebia::mente)f(\overline{prebia::mente}) - f(\overline{prebia}::mente)f(prebia::\overline{mente}))^2}{f(prebia::)f(\overline{prebia}::)f(::mente)f(::\overline{mente})}$$

$$X^2 = \frac{1{,}952{,}247\times(58\times 1.939{,}963 - 12.190\times 36)^2}{94\times 1{,}952{,}153\times 12{,}248\times 1.939{,}999}$$

$$= \frac{1{,}952{,}247\times(112{,}517{,}854 - 438{,}840)^2}{4.360219871451\times 10^{18}} = 5.624.384174299$$

También la de la razón de semejanza:

$$\begin{aligned} -2log\lambda = 2\Bigg[& logL\left(\frac{f(prebia::mente)}{f(::mente)}, f(prebia::mente), f(::mente)\right) \\ & + logL\left(\frac{f(prebia::\overline{mente})}{f(::\overline{mente})}, f(prebia::\overline{mente}), f(::\overline{mente})\right) \\ & - \left(logL\left(\frac{f(prebia::)}{T}, f(prebia::mente), f(::mente)\right)\right. \\ & \left. + logL\left(\frac{f(prebia::)}{T}, f(prebia::\overline{mente}), f(::\overline{mente})\right)\right)\Bigg]. \end{aligned}$$

donde $logL(p, n, k) = nlog(p) + (k - n)log(1 - p)$. De esta manera:

$$logL\left(\frac{58}{12248}, 58, 12248\right) = logL(0.00473547, 58, 12248) = -368.318$$

$$logL\left(\frac{36}{1939999}, 36, 1939999\right) = logL(0.0000185567, 36, 1939999) = -428.208$$

$$logL\left(\frac{94}{1952247}, 58, 12248\right) = logL(0.0000481496, 58, 12248) = -577.176$$

$$logL\left(\frac{94}{1952247}, 36, 1939999\right) = logL(0.0000481496, 36, 1939999) = -451.294$$

y por lo tanto:

$$-2log\lambda = 2 \times (-368.318 - 428.208 + 577.176 + 451.294) = 2 \times 231.944 = 463.888$$

Para el coeficiente de coligación de Yule,

$$Y = \frac{\sqrt{a}-1}{\sqrt{a}+1}, a = \frac{f(prebia::mente)f(\overline{prebia::mente})}{f(prebia::\overline{mente})f(\overline{prebia}::mente)}$$

por lo que

$$\sqrt{a} = \sqrt{\frac{58 \times 1.939,963}{12,190 \times 36}} = \sqrt{\frac{112,517,854}{438,840}}$$
$$= \sqrt{256.3983547534} = 16.01244374708$$

y

$$Y = \frac{16.01244374708 - 1}{16.01244374708 + 1} = 0.882438$$

La información mutua se calcula con

$$I = log\left(\frac{f(prebia::mente)/T}{(f(::mente)/T)(f(prebia::)/T)}\right) = log\left(\frac{Tf(prebia::mente)}{f(::mente)f(prebia::)}\right).$$
$$I = log\left(\frac{1,952,247 \times 58}{12,248 \times 94}\right) = log\left(\frac{113,230,320}{1,151,312}\right) = log(98.34894971997) = 4.588522$$

Por último, se pueden calcular estas medidas automáticamente en cada segmentación del vocablo examinado. En la tabla 2.5 se muestran los resultados del vocablo 'previamente'. Como se puede constatar aquí, mientras menor sea la medida, menos se espera que los

Tabla 2.5: Medidas estadísticas de cada segmentación de 'previamente'.

	P	R	E	B	I	A	M	E	N	T	E
χ^2	0	0	7475	33430	3622	5624	0	0	0	0	
r.s.	0	0	493.4	637.9	383.7	463.9	0	0	0	0	
Yule	0	0	0.8934	0.9303	0.8039	0.8824	0	0	0	0	
i.m.	0	0	4.87	6.36	4.162	4.589	0	0	0	0	

segmentos estén asociados y, por lo tanto, más cabe suponer allí una frontera morfológica.

Tabla 2.6: Medidas estadísticas de cada segmentación de 'aparecer'.

	A	P	A	R	E	S	E	R
χ^2		87.33	4564	10370	5904	1437	595.7	127.3
r.s.		58.25	384.5	465.6	402.5	259.7	180.3	74.25
Yule		0.3344	0.8316	0.8796	0.8373	0.7001	0.5867	0.3687
i.m.		1.161	4.476	5.285	4.73	3.367	2.569	1.399

2.4 Comparación de índices

En esta sección, las técnicas examinadas en los apartados anteriores se aplican a una muestra aleatoria de vocablos del *CEMC* con el objeto de comparar su utilidad como criterios de segmentación morfológica. Se construyó una rutina para seleccionar 851 vocablos[35] de más de cinco letras de longitud cada uno. Se calcularon los índices para cada una de sus segmentaciones y, para cada palabra, se almacenaron las segmentaciones cuyos valores fueron los mejores (los más altos —para el número de cuadros, el cociente de de Kock y la entropía— y los más bajos —para las estadísticas[36] de digramas)[37]. Además, se determinó mediante un proceso de inspección (o sea no automático) si la segmentación propuesta de cada índice (la mejor para cada vocablo) era o no válida. Para simplificar el experimento se contaron solamente los sufijos[38].

---

[35]Después de eliminar algunos como 'konboi', 'nokaut', 'juebes', etc. cuya segmentación en español es cuestionable.

[36]Para los índices estadísticos la mejor corresponde al menor valor (menos asociación) y para los otros al mayor valor (más cuadros, más economía, más información asociada a la base).

[37]De esta manera se evitó el problema de determinar un valor que sirva de umbral entre las segmentaciones morfológicamente correctas y las incorrectas. Sin embargo, esto nos dejó solamente con la posibilidad de calcular la precisión del método como medida de evaluación.

[38]Como se apuntó antes, las estadísticas de digramas no nos informan por sí mismas sobre el tipo de afijo involucrado en una segmentación. De todas maneras, estas cuatro estadísticas tienden a señalar fronteras entre bases y sufijos: sólo 31 aciertos —0.037% de la muestra—involucraban prefijos. Así, las predicciones acertadas de estas estadísticas en cuanto a las fronteras con prefijos no se tomaron en cuenta. De manera similar, se ignoraron los aciertos en la predicción de prefijos mediante los índices de cuadros, de de Kock y entropía.

Como era de esperarse, determinar si eran correctas las segmentaciones relacionadas con afijos de flexión no presentó mayores dificultades. Sin embargo, esto no fue tan sencillo con muchas que involucraban afijos de derivación. Por ejemplo, algunos vocablos aparentan tener o exhiben una estructura que no se corresponde con su significado ni con su etimología: el vocablo 'almohada' ([almoáda] > árabe: al-muhádda), para el que se propone una frontera entre almoh~ y el sufijo ~ada. Por muy etimológicamente incorrecto que sea esto, la entropía propone este análisis consistentemente y, aunque en este trabajo se contó como incorrecta, sería interesante ver qué segmentación proponen varios hablantes del español que desconozcan esta etimología y cómo saben que allí no ocurre el sufijo participial ~ada.

La mayoría de los *errores* tuvieron que ver con segmentaciones propuestas en medio de raíces muy conocidas (por ej. **su::éltame*) o de afijos también de sobra conocidos (por ej. **ekibokasióne::s*). Además, hay que observar que, aunque la mejor segmentación propuesta por un índice esté equivocada, la segunda o tercera mejores bien pueden segmentar el vocablo correctamente (la mayoría de las palabras tienen varias segmentaciones correctas). De todas maneras, en este experimento sólo se tomaron en cuenta las mejores propuestas de segmentación de cada índice.

Tabla 2.7: Comparación de índices: segmentaciones correctas en una muestra de 836 vocablos.

índice	aciertos	porcentaje
cuadros	737	87.22%
entropía	730	86.39%
de Kock-Bossaert	669	79.17%
coef. de Yule	609	72.07%
info. mutua	604	71.48%
prueba de χ^2	583	68.99%
razón de semejanza	582	68.88%
subs. entrop.	272	32.19%

En la tabla 2.7 se exhiben los resultados de esta comparación. Los vocablos de la muestra

aleatoria y los análisis de éstos se encuentran en la tabla B.4 del apéndice ("Muestra aletoria de vocablos analizados"). No es de sorprenderse que los mejores resultados los hayan obtenido los índices de cuadros, la entropía y el cociente de de Kock. los cuales parecen ser los mejores indicadores de segmentación morfológica. Hacia abajo en la tabla aparecen las medidas estadísticas de digramas y, como se dijo en la sección de entropía. la peor de las medidas es la que toma en cuenta la diferencia entre las entropías de ambas direcciones (subs. entrop.. con sólo 32% de aciertos), esto es, la substracción de los valores de entropía con respecto a los sucesores menos aquellos con respecto a los predecesores.

Un aspecto interesante del experimento es que al combinar los resultados de los índices de economía y de entropía, el porcentaje de aciertos mejoró considerablemente al 95.5%. cosa que no es sorprendente, ya que en los cuadros mismos parecen manifestarse las relaciones de economía y entropía entre los segmentos de las palabras.

La razón por la que los cuadros, la entropía y los cocientes de de Kock obtuvieron mejores resultados debe estar relacionada con, por un lado, el hecho de que se trate de medidas de propiedades que le hemos asociado al fenómeno de afijación y que toman en cuenta la estructura subyacente del inventario completo de vocablos, mientras que, por el otro. las estadísticas de digramas son medidas más generales de asociación (o no asociación) que se calculan a partir de una tabla de contingencia que no refleja la estructura lingüística del vocabulario. Es decir, los primeros se basan en conocimientos lingüísticos. Además. si los primeros son medidas de características reales de los afijos. surge la pregunta de si se pueden combinar de alguna manera para medir la propiedad. hasta ahora cualitativa. que tienen

algunos segmentos de ser afijos[39].

2.5 El catálogo de afijos

Este apartado trata de la aplicación de las técnicas expuestas y comparadas arriba en la construcción de dos catálogos de segmentos de palabras aptos de funcionar como afijos, uno para sufijos y otro para prefijos.

Aunque se ha insistido todo el tiempo que mejor o más económica será una segmentación mientras más cuadros tenga y mayor sea la diferencia entre el número de formas del supuesto afijo y de la supuesta base, no se han definido con precisión las nociones de '*suficientes* cuadros', 'formas *frecuentes*', '*más* formas', '*mayor* diferencia', etc. Es evidente que hay cierto grado de arbitrariedad en decidir cuánto es *mucho* o cuánto es *poco*, especialmente al examinar vocablos separados.

La estrategia escogida en este experimento es la de posponer estas decisiones lo más posible. Así, en lugar de rechazar el carácter de afijo de tal o cual segmento solamente con base en los datos calculados en un vocablo, es mejor hacerlo con base en todas sus ocurrencias en los vocablos del conjunto V en que aparezca, sin importar qué tan *malos* sean esos datos. Es decir, si tal o cual segmento verdaderamente representa o no un afijo, esto se verá necesariamente en el promedio de los datos calculados en cada una de sus ocurrencias.

De todos modos son necesarias restricciones mínimas, como la de requerir por lo menos

[39]En un experimento previo con una muestra menor —de 217 formas— y otra implementación —ligeramente diferente— de los índices, se obtuvieron resultados similares: las estadísticas de digramas obtuvieron entre 10 y 30 puntos porcentuales menos que las otras.

un cuadro y la de asegurarse de que el segmento examinado como afijo sea más frecuente que la supuesta base (así sea sólo por una forma). De esta manera, se puede examinar cada segmentación de cada vocablo del conjunto V y, según las medidas calculadas, insertar el segmento pertinente en alguna estructura para almacenar posibles afijos. Esta estructura sería un *catálogo* de afijos y puede definirse de la siguiente manera:

2.5.1 Definición formal

En este subapartado se define una estructura formal para almacenar los afijos a ser descubiertos mediante los criterios expuestos arriba aplicados al *Corpus del Español Mexicano Contemporáneo.*

Sea Υ un catálogo de γ afijos descrito por el séxtuplo $\langle S, C, K, H, F', F'' \rangle$, donde S es el conjunto de segmentos (o cadenas de caracteres) que aparecen como afijos, $\{s_1, s_2, s_3, \ldots, s_\gamma\}$, en el corpus Φ: C es el conjunto de promedios de cantidades de cuadros asociadas a cada ocurrencia de estos segmentos, $\{\bar{c}_1, \bar{c}_2, \bar{c}_3, \ldots, \bar{c}_\gamma\}$; K es el conjunto de promedios de índices de economía, $\{\bar{k}_1, \bar{k}_2, \bar{k}_3, \ldots, \bar{k}_\gamma\}$; H el conjunto de promedios de entropía, $\{\bar{h}_1, \bar{h}_2, \bar{h}_3, \ldots, \bar{h}_\gamma\}$; F' el de las frecuencias absolutas de cada segmento $\{\Omega'_1, \Omega'_2, \Omega'_3, \ldots, \Omega'_\gamma\}$; y F'' el conjunto de frecuencias de ocurrencia de los segmentos como afijos entre los vocablos $\{\Omega''_1, \Omega''_2, \Omega''_3, \ldots, \Omega''_\gamma\}$. De esta manera, Υ también puede describirse como el conjunto de γ relaciones ordenadas:

$$\Upsilon = \{\langle s_1, \bar{c}_1, \bar{k}_1, \bar{h}_1, \Omega'_1, \Omega''_1 \rangle,$$

$$\langle s_2, \bar{c}_2, \bar{k}_2, \bar{h}_2, \Omega'_2, \Omega''_2 \rangle,$$

$$\langle s_3, \bar{c}_3, \bar{k}_3, \bar{h}_3, \Omega'_3, \Omega''_3 \rangle,$$

$$\ldots \langle s_\gamma, \bar{c}_\gamma, \bar{k}_\gamma, \bar{h}_\gamma, \Omega'_\gamma, \Omega''_\gamma \rangle\}$$

Además, sean dos catálogos separados, Υ^p que contiene los prefijos y Υ^s que contiene los sufijos.

2.5.2 Probabilidades

En esta subsección se describen dos tipos de frecuencias relativas o probabilidades asociadas a los afijos con respecto a su pertenencia a un catálogo formal como el definido arriba.

La ocurrencia de una cadena de caracteres en un corpus no garantiza de ninguna manera que esa cadena represente a un afijo (por ej. la cadena *mente* no representa al afijo ~mente ni al pronombre sufijado ~te en los vocablos 'comente', 'aumente', 'argumente', etc.). Como hemos visto, los índices presentados arriba permiten determinar con cierta seguridad cuándo un segmento es un afijo y cuándo no.

En la teoría clásica de la probabilidad[40], las probabilidades se pueden estimar directamente de los datos recolectados empíricamente simplemente al dividir el número de ocurrencias de un evento entre la suma de las ocurrencias del total de los eventos. Así, podemos calcular la probabilidad de aparición de un vocablo v_i en el corpus Ψ de la siguiente manera:

$$0 \leq p(v_i) = \frac{f_i}{\xi} \leq 1, \quad i = 1, 2, 3, ..., \Omega.$$

donde, como quedó establecido arriba, f_i es la frecuencia del vocablo v_i, ξ es el tamaño del corpus Ψ y Ω es el total de vocablos (el tamaño del conjunto V).

De manera similar, al contar el número de veces que un segmento cumple su papel de afijo,

[40]Para una presentación de esta teoría en el marco de la lingüística, véanse Piotrowski, Bektaev y Piotrowskaja, *Mathematische Linguistik* [113], Brockmeyer, Bochum, 1985, pp. 153-165, y Piotrowski, Lesohin y Lukjanenkov, *Introduction of Elements of Mathematics to Linguistics* [114], Brockmeyer, Bochum, 1990, pp. 162-216.

se puede calcular la probabilidad de que el segmento sea un afijo (al escoger aleatoriamente un vocablo del conjunto V que contenga el segmento). Si contamos en el catálogo Υ con un conjunto $F' = \{\Omega'_1, \Omega'_2, \Omega'_3, \ldots, \Omega'_\gamma\}$ de las frecuencias absolutas de los segmentos y un conjunto $F'' = \{\Omega''_1, \Omega''_2, \Omega''_3, \ldots, \Omega''_\gamma\}$ de las frecuencias de cada forma como afijo ($\Omega''_k \leq \Omega'_k \leq \Omega$, para cada $k = 1, 2, 3, \ldots, \gamma$). Entonces, la probabilidad de que un segmento s_k sea un afijo se puede estimar de la siguiente manera:

$$0 \leq p(s_k) = \frac{\Omega''_k}{\Omega'_k} \leq 1, \qquad k = 1, 2, 3, \ldots, \gamma. \tag{2.5}$$

También es posible tomar en cuenta que cada vocablo en que aparece un segmento tiene su propia frecuencia de ocurrencia en el corpus. Además, un subconjunto de esas ocurrencias corresponderá a apariciones con carácter de afijo. Recuérdese que f_i es la frecuencia del vocablo i en el corpus. Si contamos en el catálogo Υ con un conjunto de segmentos S, cada s_k tendrá una $f'_{k,i}$ igual a f_i y una $f''_{k,i}$ que será igual a f_i cuando s_k sea afijo de v_i, e igual a 0 cuando no:

$$f'_{k,i} = f_i. \qquad f''_{k,i} = \begin{cases} f_i, & s_k \in A_i^p, s_k \in B_i^s \\ 0, & s_k \notin A_i^p, s_k \notin B_i^s \end{cases}$$

Así, se puede calcular una segunda probabilidad que toma en cuenta las frecuencias de los vocablos en el corpus Ψ:

$$0 \leq p^{\Psi}(s_k) = \frac{\sum_{q=1}^{\Omega''_k} f''_{k,q}}{\sum_{r=1}^{\Omega'_k} f'_{k,r}} \leq 1, \quad k = 1, 2, 3, \ldots, \gamma. \tag{2.6}$$

Esto quiere decir que al recorrer una cadena de palabras Ψ, $p^{\Psi}(s_k)$ es la probabilidad de que el segmento s_k de la última palabra examinada sea un afijo.

Estas probabilidades se asemejan a las sugeridas por Meya para medir la probabilidad de ocurrencia de un morfema de alguna lengua a partir de una cadena de Markov (cosa, como establece la investigadora, muy útil cuando menos en procedimientos de síntesis automática del habla, ya que las fronteras de morfemas proporcionan información sobre la prosodia[41]). No es difícil imaginar cómo las probabilidades arriba descritas se pueden afinar mediante la construcción de cadenas de Markov (véase el capítulo siguiente para la definición, discusión y aplicación del esquema markoviano). La diferencia principal es que Meya presupone las reglas morfológicas (su red de hipótesis específicas al español y confeccionadas manualmente) para segmentar palabras en morfemas, mientras que aquí se propone un esquema de segmentación que presupone cuando mucho una estructura morfológica descriptible mediante índices de entropía, economía y cuadros.

2.6 Hacia un índice de *afijalidad*

En este apartado se investiga la manera de calcular un índice general de *afijalidad*, es decir, una medida del carácter de afijo que pueda tener un segmento de vocablo dado. La notable ventaja de las medidas de economía, entropía y número de cuadros en la predicción de fronteras morfológicas tal vez pueda explicarse en el hecho de que éstas corresponden a nuestra concepción tradicional de cómo son los afijos (son menos, más frecuentes y contienen menos información que otros tipos de signos). Por lo tanto, tiene sentido intentar caracterizar formalmente la cualidad que un segmento de palabra pueda tener de ser un afijo en términos de por lo menos estos tres índices.

[41] Meya, art. cit. [102] 1986, p. 142.

En la introducción me referí a esta cualidad como la *afijalidad* de un segmento s_x y, adelantando los buenos resultados de la entropía, economía y cuadros, se propuso una hipótesis para medirla:

$$AF(s_x) = \frac{f_x c_x k_x}{h_x} \tag{2.7}$$

donde f_x, c_x, k_x y h_x representan respectivamente la frecuencia, la cuenta de cuadros, la economía y la entropía del segmento s_x calculadas a partir de los vocablos de un corpus Ψ. Como ya se explicó desde la introducción, para que un segmento s_x sea un afijo, se espera, por un lado, que ocurra con una gran frecuencia en el corpus, que esté involucrado en un gran número de cuadros c_x, y que tenga asociada una gran cantidad de economía k_x. Por el otro, un afijo deberá contener una cantidad mínima de información h_x. En la fórmula 2.7 mientras mayores sean las primeras cantidades y menor sea la cantidad de información, mucho mayor será $AF(s_x)$.

Sin embargo, hay dos cosas que hay que reconsiderar sobre esta fórmula. Primero, es obvio que la frecuencia del segmento es una señal de su afijalidad, pero también es cierto que en la frecuencia se manifiestan fluctuaciones de diversos tipos (de hecho, la tendencia de ciertos afijos a ocurrir en ciertos tipos de textos y no en otros no es un problema trivial). Es más, la frecuencia misma puede concebirse como una manifestación o resultado de la economía, las estructuras combinatorias y el contenido de información inherentes al afijo. De hecho, estos índices parecen lo suficientemente robustos como para rendir cuenta del fenómeno que nos interesa.

Segundo, como se mencionó arriba, el mínimo contenido de información como criterio para determinar afijos no funciona tan bien como el pico de mayor información que caracteriza a

las bases. En esencia se trata del hecho de que los afijos no son los segmentos con menor información, sólo contienen menos que las bases. Por eso, vale la pena tomar en cuenta no la entropía que *contiene* un afijo, sino la entropía de lo que lo rodea, que es una marca de frontera entre base y afijo y es directamente proporcional a la afijalidad.

Al tomar estas consideraciones en cuenta, podemos redefinir a la afijalidad de la siguiente manera:

$$AF(s_x) = k_x c_x h_x$$

Esto es, la cualidad que tiene s_x de ser un afijo es directamente proporcional al producto de alguna medida de economía (k) por el número de cuadros (c), por una medida (h) de la sorpresa inherente a la transición de ese segmento al siguiente —todas estas cantidades calculadas a partir de la frontera de un segmento y sus posibles segmentos adyacentes (supuestas bases) y del conjunto V de vocablos.

Esta generalización es válida para el cálculo de la afijalidad de segmentos de vocablos aislados (afijos-ocurrencia), pero la misma relación se sostiene para los promedios de los valores de afijalidad de varios afijos-ficha, es decir, de los afijos-tipo, cosa que resulta en un índice de afijalidad para las formas que pertenecen al catálogo de afijos Υ:

$$AF(s_x) = \bar{k}_x \bar{c}_x \bar{h}_x$$

De manera intuitiva, estos tres índices se pueden combinar para representar otros tipos de fenómenos morfológicos. Por ejemplo, sería interesante examinar detenidamente la relación entre el cociente de de Kock-Bossaert (como medida de economía) y el número de cuadros

—el cual, por sí mismo, parece medir qué tanto se usa el segmento como afijo[42]. Así, un índice bajo de economía, un número pequeño de cuadros o poca entropía disminuirían la afijalidad del segmento. Pero el que uno de estos índices desfavorezca la afijalidad general, no significa que alguno de los otros dos —o los dos— no se deban tomar en cuenta: puede tratarse de indicadores de algún otro tipo de morfo. De hecho, sería pertinente explorar si —y hasta qué punto— un índice bajo de economía junto a un número alto de cuadros estarían relacionados al fenómeno de composición. De manera similar, una medida baja de información seguramente estaría relacionada con algún tipo de unidad morfológica de contenido, porque la medida no mide su contenido de información, sino el de lo que le es adyacente según el corpus[43].

Pero aparte de las virtudes de combinar estas medidas, hay que tomar en cuenta la cuestión de la normalización (esto es, en el intervalo [0,1]). Hay varias maneras de proceder para normalizar nuestra definición de afijalidad. Una posibilidad es la fórmula siguiente:

$$AF^n(s_x) = 1 - \frac{1}{AF(s_x)} \leq 1$$

que funciona muy bien con los datos extraídos en este experimento[44]. Pero hay otras posibilidades de normalización. Sin embargo, para evitar las preguntas que surgen acerca de la

[42]La aplicación del término *productividad* es tentadora, pero su uso aquí sería cuestionable, ya que dicho término ya se ha usado para designar ideas más complejas; véase, por ejemplo, en el marco de la lingüística cuantitativa, el trabajo de Baayen, que ha investigado a profundidad las maneras de medir la productividad de reglas morfológicas a partir de frecuencias de vocablos (*types*) y palabras (*tokens*), Baayen, *op. cit.* [12] 1989, especialmente el capítulo 2, pp. 27-54.

[43]Recuérdese que en este experimento, la medida de entropía se refiere a la transición entre un supuesto afijo y sus bases potenciales (así, aunque el afijo contiene menos información, la entropía que aquí se le asoció es en realidad el contenido de información de la base).

[44]Así, en el catálogo de presuntos afijos compilado automáticamente (véase descripción abajo), esta fórmula le asignó a unos pocos segmentos la puntuación perfecta de 1 (los *mejores*, casi todos de flexión). La gran mayoría recibió un valor dentro del intervalo [0,1] y a los últimos se les asignó un valor negativo. Al inspeccionar estos últimos —y como era de esperarse—, no hubo ningún afijo.

naturaleza de las magnitudes y las escalas de los índices a combinar, la estrategia escogida para este experimento fue el cálculo del promedio de los índices normalizados: cada promedio de cada índice fue normalizado dividiéndolo por el valor máximo obtenido para ese índice (cuando no se conocía el máximo global, se utilizó el máximo del vocablo). Luego, para evitar números demasiado pequeños, se sacó el promedio de los índices normalizados, en lugar de multiplicarlos:

$$AF^n(s_x) = \frac{\frac{c_x}{\max c_i} + \frac{h_x}{\max h_i} + \frac{k_x}{\max k_i}}{3} \tag{2.8}$$

Así, este índice de afijalidad se calculó para los vocablos de la muestra aleatoria utilizada en la comparación de índices presentada arriba. 764 de los 836 vocablos fueron segmentados correctamente, lo que significó el 90.41% (es de notarse que este porcentaje es menor que el índice calculado sin las cuentas de cuadros que, como se dijo arriba, obtuvo el 95.5% de aciertos). En las tablas 2.12 y 2.13 (páginas 138 y 140, respectivamente), la afijalidad se calculó tomando los tres índices en cuenta siguiendo una estrategia similar a la de la fórmula 2.8, pero usando los promedios de cada uno (cada promedio de cada índice fue normalizado y luego promediado con los otros índices de la segmentación).

2.7 Catálogos de afijos a partir del *CEMC*

Esta sección representa la aplicación de lo presentado en las secciones anteriores. Concretamente, se describen los resultados de la construcción de un catálogo de sufijos y otro de prefijos del español.

Para construir los catálogos Υ^s y Υ^p a partir del *CEMC*, se llevaron a cabo varios experimentos que permitieron explorar los diferentes caminos posibles en la recolección de los

afijos. Un aspecto importante de esto fue la adaptación de los grafemas para corresponder a los fonemas. La tabla A.7 del primer apéndice resume las modificaciones de caracteres hechas a los vocablos para reflejar las particularidades del español de México.

También es importante tomar en cuenta la acentuación gráfica de las palabras. El hecho de que en el texto español se representen gráficamente las sílabas tónicas mediante un acento sobre la vocal tónica (según reglas bien conocidas) y que varios morfemas se distingan entre sí sólo por esta diferencia (esto es, $\sim o \neq \sim ó$, $\sim aras \neq \sim arás$, and $\sim ás \neq \sim as$) es una invitación a mantener los acentos de las palabras gráficamente acentuadas. El problema es la introducción de acentos en las sílabas tónicas sin acento gráfico, específicamente, en las palabras graves que terminan en 'n', 's' ni vocal. Esto es un inconveniente porque requiere presuponer la estructura de la sílaba[45]. La solución parcial consistió en conservar solamente los acentos de las últimas sílabas (es decir, de las últimas vocales). A las palabras agudas que en su escritura terminaban en otra cosa que no fuera 'n', 's' o vocal se les agregó el acento gráfico a la última vocal: 'coronel' → [koronél], 'vejez' → [bejés], 'codorniz' → [kodornís].

El paso siguiente fue examinar automáticamente cada posible segmentación j de cada vocablo v_i para determinar —también automáticamente— sus índices pertinentes: $c^p_{i,j}$ (número de cuadros al asumir un prefijo), $c^s_{i,j}$ (número de cuadros al asumir un sufijo)), $k^p_{i,j}$ (índice de de Kock al asumir prefijo), $k^s_{i,j}$ (índice de de Kock al asumir sufijo), $h^p_{i,j}$ (entropía de prefijo a base), $h^s_{i,j}$ (entropía de sufijo a base), etc. que nos proporcionan los criterios para determinar qué tan morfológica es la segmentación. De estos. se calcularon para cada segmentación dos índices de afijalidad: $AF^p_{i,j}$ y $AF^s_{i,j}$ (uno hipotetizando un prefijo y el otro suponiendo un

[45]Como se mencionó en la introducción, esto iría en contra de la idea de tomar al corpus como el dato, sin agregarle nada manualmente, por obvio que sea.

sufijo).

Tabla 2.8: Medidas de segmentación del vocablo 'aumente'.

	A	U	M	E	N	T	E
cuadros, c^s	0		1021	20	0	0	**8348**
entropía, h^s	0		1.046	1.066	0.6184	1.351	**2.018**
de Kock, k^s	0		0.9207	0.55	0	0	**0.9251**
afijalidad, AF^s	0		0.5453	0.3752	0.1022	0.2231	**1**

Tabla 2.9: Medidas de segmentación del vocablo 'comente'.

	C	O	M	E	N	T	E
cuadros, c^s	0		0	0	1009	602	**6505**
entropía, h^s	1.099		1.046	1.066	0.6184	1.351	**2.018**
de Kock, k^s	0		0	0	0.4757	0.9402	**0.9445**
afijalidad, AF^s	0.1815		0.1728	0.1762	0.3218	0.5858	**1**

Tabla 2.10: Medidas de segmentación del vocablo 'previamente'.

	P	R	E	B	I	A	M	E	N	T	E
cuadros, c^s	0	0	0	**6**	**1050**	468	0	0	0	0	
entropía, h^s	0	0	1.609	1.022	**2.233**	1.046	1.066	0.6184	1.351	2.018	
de Kock, k^s	0	0	0	0	0.9657	**0.9915**	0	0	0	0	
afijalidad, AF^s	0	0	0.2402	0.1525	**0.9913**	0.638	0.1592	0.0923	0.2016	0.3011	

Compárense las segmentaciones posibles de los vocablos en las tablas 2.8, 2.9 y 2.10. Cabría esperar que la alta frecuencia del sufijo *~mente* señalara cortes incorrectos al interior de raíces tales como *aument~* y *coment~*, pero las mejores segmentaciones según todos los índices para ambos vocablos 'comente' y 'aumente' proponen al sufijo flexional *~e*. En ambas tablas tenemos una puntuación perfecta de afijalidad porque para normalizar los valores se tomó en cuenta el valor máximo de cada palabra. Nótese además, cómo *~mente* no es el único segmento compitiendo como afijo (el segmento *~te* que se asemeja al pronombre enclítico que normalmente acompaña gerundios, infinitivos e imperativos también obtuvo un valor alto —mayor de 0.5— en el vocablo 'comente'). Obsérvese también cómo en la presencia del sufijo *~mente* (tabla 2.10), otras segmentaciones pueden exhibir una afijalidad más alta. Así, aunque *~mente* en 'previamente' corresponde a la medida más alta del índice de de Kock-

Tabla 2.11: Medidas de segmentación del vocablo 'nacionalidad'.

	N	A	S	I	O	N	A	L	I	D	A	D
c^s	0	0	0	0	0	*15*	0	**135**	1	0	0	
h^s	0.5623	1.792	0	0.6277	0.9592	**2.133**	0.9149	*1.782*	0.5373	0.3152	0.8632	
k^s	0	0	0	0	0	*0.4*	0	**0.9333**	0	0	0	
AF^s	0.08789	0.28	0	0.0981	0.1499	*0.5132*	0.143	**0.9451**	0.08398	0.04926	0.1349	

Bossaert, el mejor valor de afijalidad favorece la secuencia de sufijos *~a-mente*. De todas maneras, *~mente* obtiene el segundo valor más alto, un muy respetable 0.638.

Como exhiben estos ejemplos, es conveniente determinar un valor mínimo que sirva de umbral para evitar la aceptación de segmentos con una afijalidad demasiado baja. En la compilación de los catálogos, se consideró suficiente aceptar los segmentos con 0.5 o más para los sufijos, mientras que se requirió un valor igual o mayor de 0.8 para los prefijos. Cada vez que se obtuvieron estos valores mínimos, se insertó el segmento en cuestión en el catálogo apropiado (si un prefijo, el segmento izquierdo en Υ^p y, si un sufijo, el derecho en Υ^s). Si dicho segmento ya estaba presente en el catálogo pertinente, se pusieron al día los registros correspondientes a cada índice.

Como ya se dijo arriba, muchos vocablos contienen más de un afijo. De allí que fuera tan común encontrar más de una segmentación válida. Con respecto a esto, hay varias alternativas posibles a seguir, según se acepten una o varias de las segmentaciones:

1. tomar todos los segmentos con afijalidad > 0,
2. tomar solamente el segmento con el mejor índice de afijalidad (*~amente* en la tabla 2.10 e *~idad* en la tabla 2.11) o,
3. tomar todos los segmentos con índices mayores a un valor umbral (al requerir 0.5 de afijalidad, se aceptarían *~amente* y *~mente* en la misma tabla) o,
4. aplicar algún algoritmo, por ejemplo:

(a) tomar el segmento con la afijalidad más alta y, recursivamente, tomar el próximo más alto entre la segmentación aceptada y la raíz del vocablo (hacia la izquierda cuando se trate de un sufijo); esto es, seleccionar el mejor, luego ignorar los valores al interior del supuesto afijo (o cadena de afijos) que colinda con ese corte y buscar el siguiente mejor valor en lo que queda del vocablo (en la tabla 2.11, tomar el afijo *~idad* y luego *~alidad*),

(b) tomar el segmento con la afijalidad más alta y, recursivamente, tomar el próximo más alto entre la segmentación aceptada y el exterior del vocablo (hacia la derecha cuando se trate de un sufijo); esto es, seleccionar el mejor, luego buscar los cortes entre éste y el exterior del vocablo (en la tabla 2.10, tomar *~amente* y luego *~mente*),

(c) etc.

La alternativa depende del tipo de catálogo que se quiera conseguir. Por ejemplo, para un conjunto pequeño de afijos de flexión podrán seleccionarse los afijos (o cadenas de afijos) entre el mejor corte y el exterior del vocablo (algoritmo b). La primera alternativa (tomar todos los valores > 0) necesariamente resultaría en el catálogo más grande posible, que contendría el número mayor de afijos falsos. Como se dijo arriba, en este proyecto se tomó la segunda alternativa. Sin embargo, se mantuvieron cuentas para registrar si los segmentos correspondían a las mejores segmentaciones, a las cadenas de caracteres más exteriores o más interiores.

Una vez que fueron examinadas automáticamente todas y cada una de las segmentaciones de cada vocablo del conjunto V y los segmentos pertinentes insertados en los catálogos Υ^s y Υ^p, se observaron varias cosas que pudieron haberse esperado. Muchas de las cadenas de caracteres que se recolectaron son grupos de afijos tanto de derivación como de flexión. Los pocos segmentos con los promedios más altos de afijalidad son afijos o grupos de afijos españoles de sobra conocidos. En cambio, la mayoría son de frecuencia relativamente muy baja y tienen promedios de afijalidad más bien bajos: mientras menor sea el índice de afi-

jalidad, más dudas surgen en cuanto a la cualidad morfológica de la cadena de caracteres en cuestión. Las fronteras entre afijos bien conocidos, afijos dudosos y residuos obviamente no morfológicos no son claras. De hecho, algunos afijos conocidos no pudieron distinguirse formalmente de los *errores* obvios mediante el procedimiento aquí aplicado. Por ejemplo, el sufijo de flexión verbal *~áis*, que en México no es muy productivo, obtuvo el rango 2085 con un índice de afijalidad de 0.3198 y aparece muy cerca de afijos falsos como *~gún* de 'algún' o 'ningún'.

La curva en la figura 2.3 muestra los valores de afijalidad de todos los segmentos acumulados en el catálogo Υ^s en orden de mayor a menor (incluidos los afijos *falsos* que, como se dijo arriba, tienen valores menores y, por lo tanto, aparecen hacia la derecha de la curva. De manera similar, la figura 2.4 muestra la curva de valores de afijalidad de los prefijos en orden de rango. Como puede verse, la diferencia principal está en el número de segmentos. A ambos grupos de datos se les aplicó el ajustador de Altmann (Altmann-Fitter[46]) y se observó que ambos se ajustan a la distribución hipergeométrica negativa.

Las frecuencias registradas permiten el cálculo de las probabilidades descritas arriba. Sin embargo, la relación entre éstas y las afijalidades de los segmentos no son lo que uno se hubiera esperado en un principio: los valores mayores para cualquiera de los tipos de probabilidad no necesariamente implican un mayor índice de afijalidad y los valores menores no corresponden a índices. Las mayores probabilidades sólo implican una mayor certidumbre de afijalidad, lo que de ninguna manera significa una mayor afijalidad del segmento en cuestión.

Por ejemplo, el sufijo *~a* —que obtuvo un altísimo índice de afijalidad, pero bajas pro-

[46] *Iterative Fitting of Probability Distributions* [5], Lüdenscheid: RAM, 1997.

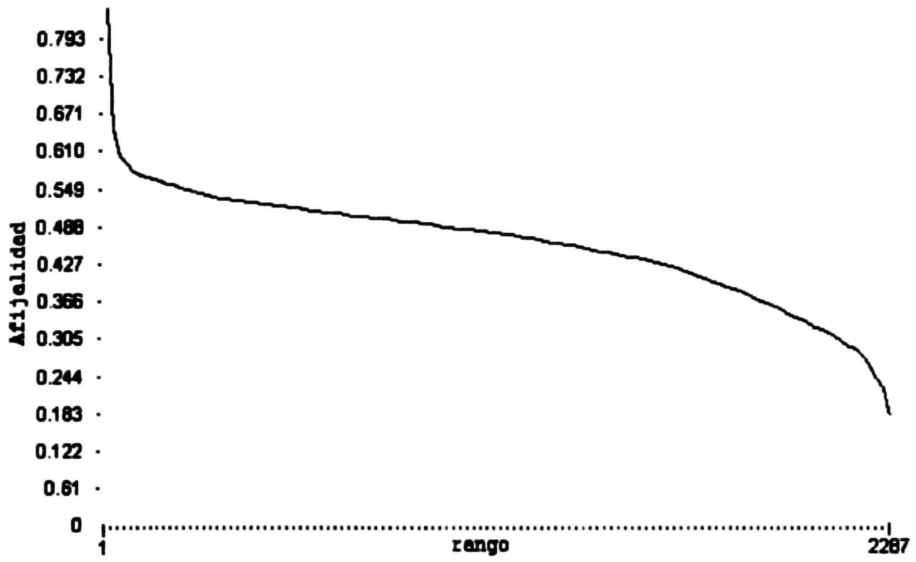

Figura 2.3: Distribución de los valores de afijalidad (sufijalidad) de todos los segmentos recogidos en el catálogo de sufijos del español de México.

babilidades (véase la tabla 2.12)— no fue considerado tal en segmentos como *~aba*, *~iera* o *~ería* de vocablos como 'compraba', 'sufriera' y 'tortillería'. Así, sus posibilidades de aparecer como sufijo resultaron mucho menores que aquellas de los sufijos largos como *~mente* o secuencias de sufijos como *~ándoselas*. Considérese también el sufijo *~ó*, que obtuvo probabilidades más bien altas. Esto significa que al encontrarnos casualmente con el vocablo 'buscó', podemos esperar que *~ó* sea un sufijo con más seguridad que si nos encontráramos con 'busca' y esperáramos que *~a* fuera un afijo también. De esta manera, al observar una cadena de ocurrencias de palabras en contexto, tal vez podamos utilizar estas probabilidades como criterios de confianza en la determinación de morfemas.

Nótese también cómo mientras más largos son los segmentos, mayor es la probabilidad de que sean afijos o una secuencia de ellos, mientras que los sufijos aislados de flexión tienden a tener probabilidades menores.

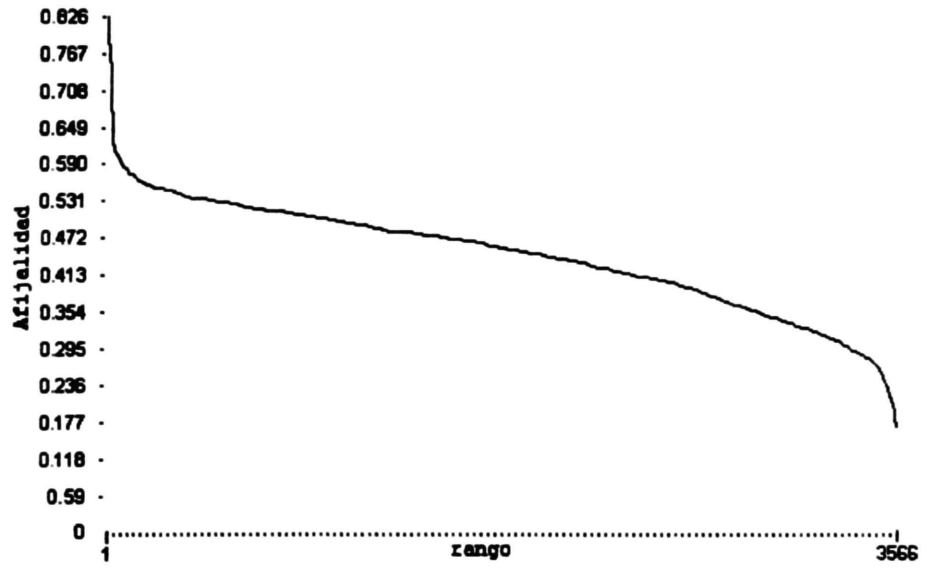

Figura 2.4: Distribución de los valores de afijalidad (prefijalidad) de todos los segmentos recogidos en el catálogo de prefijos del español de México.

Además, obsérvense las diferencias entre los dos tipos de probabilidad. Mientras mayor es la segunda probabilidad con respecto a la primera, más frecuentes son las formas en las que el afijo aparece. Esto quiere decir sencillamente que el afijo es muy frecuente en el corpus (por ejemplo, en la tabla 2.12: *~ado*, *~ada*, *~asión*, *~asiones*, *~amiento*; en el catálogo de sufijos completo, no hubo afijos exclusivamente de flexión con una probabilidad del segundo tipo mayor que la del primero con una diferencia de más de 0.2). Por otra parte, si la primera probabilidad es mayor que la segunda, el afijo en cuestión ocurre en formas poco frecuentes. Así, los pronombres enclíticos que aparecen en gerundios, infinitivos e imperativos pertenecen a este grupo: *~lo*, *~la*, *~se*, *~los*, *~las* tienen una primera probabilidad más alta que la segunda con una diferencia de al menos 0.4; *~le* y *~me* con una diferencia de al menos 0.3; y *~nos* con una de por lo menos 0.2.

Las tablas 2.12 y 2.13 contienen una selección de cadenas de sufijos y de prefijos del *CEMC*

Tabla 2.12: Selección de sufijos del español según el *CEMC* en orden de *afijalidad*

	sufijo	fr.[a]	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
1.	~ó	1428	0.7371	0.9192	0.872	0.8745	0.9003	0.8428
2.	~o	6314	0.686	0.9788	0.8017	0.4695	0.6291	0.8222
3.	~s	12013	1	0.9968	0.4609	0.5378	0.5125	0.8192
4.	~a	7687	0.5753	0.9818	0.8888	0.5153	0.4431	0.8153
5.	~os	4554	0.4775	0.9754	0.8235	0.5162	0.5639	0.7588
6.	~as	4324	0.4216	0.9779	0.8645	0.6075	0.5965	0.7547
7.	~en	945	0.4107	0.8991	0.906	0.863	0.2368	0.7386
8.	~ar	1633	0.2178	0.9621	0.9149	0.7346	0.8928	0.6982
9.	~ado	1429	0.2061	0.9619	0.907	0.7099	0.9231	0.6917
10.	~ando	976	0.1836	0.9544	0.9162	0.8399	0.9708	0.6847
11.	~e	2363	0.42	0.9482	0.6817	0.2738	0.2295	0.6833
12.	~é	639	0.4104	0.8198	0.8153	0.8925	0.409	0.6818
13.	~aba	828	0.1821	0.9565	0.9024	0.8894	0.9564	0.6803
14.	~aron	736	0.1779	0.9604	0.8935	0.8943	0.9726	0.6773
15.	~ada	1135	0.1654	0.9491	0.9159	0.7385	0.9227	0.6768
16.	~arse	665	0.1462	0.9541	0.9072	0.8428	0.9521	0.6692
17.	~ados	941	0.1477	0.9549	0.9008	0.7189	0.8582	0.6678
18.	~aban	551	0.1434	0.9395	0.9002	0.9062	0.9578	0.661
19.	~adas	813	0.1316	0.9449	0.9041	0.767	0.8687	0.6602
20.	~an	1775	0.195	0.9434	0.8354	0.6187	0.6729	0.6579
21.	~ara	370	0.1098	0.9151	0.9151	0.8916	0.9848	0.6467
22.	~ará	387	0.121	0.9295	0.8739	0.9214	0.9021	0.6415
23.	~arlo	316	0.09269	0.9291	0.8849	0.9159	0.9588	0.6356
24.	~arla	270	0.0795	0.9185	0.9071	0.931	0.965	0.635
25.	~arme	244	0.08683	0.9134	0.8916	0.9313	0.9537	0.6306

[a]Las frecuencias no se refieren a las de las formas en el corpus, sino al número de vocablos en que obtuvieron el valor de afijalidad más alto, con respecto a las otras segmentaciones posibles de cada vocablo.

(cadenas de un sólo morfo encabezan ambas tablas). En la primera aparecen las cincuenta formas con los valores máximos de afijalidad de entre 2287 del catálogo Υ^s. La mayoría son sufijos de flexión verbal e incluyen la vocal temática [*a*] (de la primera conjugación, que es la más numerosa), algunos de los cuales terminan con algún pronombre enclítico (*~te*, *~se*, *~la*, etc.). También aparecen algunos pocos sufijos exclusivamente de derivación (*~asión*, *~ador*) y a veces ocurren también sus formas en plural (*~antes* de *~ante*; *~asiones* de *~asión*).

Algunas formas se distinguen entre sí solamente mediante un acento gráfico, cosa que refleja el hecho de que correspondan a diferentes significados (*~ará*, 3ª singular del futuro: *~ara*, 1ª y 3ª singular subjuntivo pasado). Estas distinciones se habrían perdido si no se

Tabla 2.12 (continuación):
Selección de sufijos del español según el *CEMC* en orden de *afijalidad*

	sufijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
26.	~andose	260	0.07949	0.9136	0.8855	0.8966	0.9067	0.6262
27.	~arán	256	0.08995	0.9112	0.8759	0.9242	0.9118	0.6257
28.	~ido	445	0.1038	0.8567	0.914	0.7672	0.8516	0.6248
29.	~ita	453	0.09631	0.8965	0.8729	0.7639	0.8077	0.6219
30.	~aría	231	0.08063	0.8869	0.892	0.924	0.9059	0.6198
31.	~amos	645	0.1345	0.8801	0.8415	0.738	0.8804	0.6187
32.	~amente	624	0.1189	0.9784	0.7534	0.8607	0.9793	0.6169
33.	~arlos	201	0.06461	0.8959	0.8853	0.9095	0.9235	0.6153
34.	~ador	268	0.05489	0.8927	0.8965	0.7768	0.7696	0.6147
35.	~aran	196	0.07454	0.8688	0.8857	0.9245	0.9145	0.6097
36.	~es	2479	0.1876	0.9529	0.6885	0.4846	0.6193	0.6097
37.	~ito	421	0.09323	0.8752	0.8594	0.736	0.7269	0.6093
38.	~aste	136	0.05731	0.8344	0.9237	0.8662	0.8713	0.6051
39.	~arte	144	0.05526	0.8446	0.9136	0.9057	0.8802	0.6045
40.	~antes	187	0.03646	0.8802	0.8944	0.6404	0.4992	0.6037
41.	~adores	196	0.04172	0.8653	0.9029	0.7717	0.8551	0.6033
42.	~idos	269	0.07274	0.8321	0.9042	0.727	0.8084	0.603
43.	~ida	304	0.08306	0.8372	0.8864	0.7221	0.9108	0.6022
44.	~arlas	139	0.05346	0.8775	0.8755	0.891	0.9136	0.6021
45.	~asión	540	0.06338	0.9278	0.8111	0.5273	0.8398	0.6007
46.	~amiento	141	0.02085	0.8871	0.8935	0.6104	0.8256	0.6005
47.	~ante	240	0.03402	0.8769	0.8883	0.6383	0.7293	0.5998
48.	~arle	176	0.06573	0.8618	0.8716	0.9514	0.9866	0.5997
49.	~asiones	263	0.04548	0.9155	0.8365	0.6726	0.8612	0.5992
50.	~aremos	104	0.04195	0.8431	0.9103	0.9369	0.8746	0.5985

hubieran tomado en cuenta los acentos gráficos de la última sílaba[47].

En general, el catálogo de sufijos contiene una lista relativamente limpia de formas reconocibles donde los índices más bajos de afijalidad corresponden a casos dudosos, formas poco productivas y segmentos que no pueden considerarse sufijos tomando en cuenta criterios cualitativos.

Por otra parte, el catálogo de prefijos contiene más de 3566 tipos —muchos más que el de sufijos. De hecho, para limpiar la lista lo más posible de manera automática se aplicó un filtro para eliminar los segmentos que obtuvieran un valor de afijalidad menor a 0.8 (un

[47] Además, como se esperaba, la ausencia de acentos gráficos en otras sílabas no causó desajustes (*~andose* no se distingue de *~ándose*).

umbral mucho mayor que el de 0.5 aplicado a los sufijos). La explicación más inmediata de esta diferencia es el alto *ruido* que causan las múltiples y variadas abreviaturas (siglas, símbolos químicos, etc., que hubieran tenido que extraerse manualmente) en los datos en general. Tal vez se pueda recurrir a esto mismo para explicar los aparentes errores dentro los mejores 50 prefijos de la tabla 2.13 (como aparentemente serían *r̄i~* y *či~* en vocablos como 'riendas', 'riachuelo', 'chicoteo', 'chismorreo', etc.). El número de cuadros es el índice

Tabla 2.13: Selección de prefijos del español según el *CEMC* en orden de *afijalidad*

	prefijo	fr.[a]	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
1.	a~	2074	0.05189	0.9511	0.963	0.2175	0.2454	0.6553
2.	r̄e~	1866	0.07725	0.9639	0.9108	0.4999	0.6156	0.6507
3.	semi~	63	0.004754	0.943	0.9617	0.7	0.4078	0.6365
4.	des~	1041	0.05772	0.97	0.847	0.4813	0.7355	0.6249
5.	auto~	82	0.01445	0.9076	0.9144	0.4767	0.5803	0.6122
6.	pro~	481	0.02093	0.9166	0.8892	0.4546	0.7168	0.6089
7.	kontra~	92	0.01568	0.8883	0.9219	0.4532	0.1738	0.6086
8.	anti~	66	0.004184	0.9108	0.8976	0.4125	0.1791	0.6042
9.	in~	1066	0.03447	0.9551	0.8176	0.3731	0.4243	0.6024
10.	sub~	198	0.01012	0.8532	0.9334	0.6187	0.6009	0.5989
11.	r̄adio~	31	0.02116	0.8803	0.8728	0.7209	0.3871	0.5914
12.	porta~	16	0.04622	0.9062	0.8216	0.3556	0.283	0.5913
13.	sali~	15	0.06188	0.9171	0.7908	0.3488	0.6266	0.5899
14.	intra~	21	0.001767	0.9113	0.8444	0.4286	0.5	0.5858
15.	sobre~	126	0.01693	0.9298	0.8063	0.8182	0.1228	0.5844
16.	trai~	17	0.1323	0.9348	0.6688	0.3469	0.3522	0.5786
17.	idro~	32	0.002454	0.9122	0.8157	0.5246	0.6757	0.5768
18.	pre~	425	0.02103	0.9115	0.7917	0.4516	0.4459	0.5748
19.	per~	239	0.01326	0.8465	0.8512	0.3469	0.2619	0.5703
20.	r̄i~	44	0.00775	0.7299	0.9549	0.2178	0.2388	0.5642
21.	inter~	139	0.01002	0.8475	0.8276	0.5055	0.4259	0.5617
22.	kon~	866	0.02945	0.9429	0.7018	0.3991	0.3115	0.5581
23.	electro~	21	0.01301	0.8139	0.8398	0.4468	0.601	0.5556
24.	mono~	30	0.02147	0.8125	0.8317	0.5085	0.5187	0.5552
25.	dis~	217	0.01129	0.8777	0.7739	0.3344	0.2682	0.5543

[a]Las frecuencias no se refieren a las de las formas en el corpus, sino al número de vocablos en que obtuvieron el valor de afijalidad más alto, con respecto a las otras segmentaciones posibles de cada vocablo.

probablemente más afectado. Aunque no son prefijos reconocibles, muy pocas formas con muy baja afijalidad obtuvieron cuentas gigantescas de cuadros. Nótese también en la misma tabla cómo todos los mejores 50 prefijos tienen índices de cuadros muy muy bajos —mientras

Tabla 2.13 (continuación):
Selección de prefijos del español según el *CEMC* en orden de *afijalidad*

	prefijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
26.	super~	43	0.004594	0.7956	0.8605	0.3333	0.1255	0.5536
27.	pi~	165	0.0187	0.7508	0.887	0.3167	0.2764	0.5522
28.	neo~	16	0.001004	0.8473	0.8064	0.4706	0.5	0.5516
29.	psiko~	20	0.001512	0.927	0.7183	0.5128	0.5209	0.549
30.	mikro~	27	0.001804	0.8225	0.8189	0.7105	0.7653	0.5477
31.	jeo~	24	0.001396	0.9156	0.692	0.6316	0.7895	0.5363
32.	laba~	15	0.0557	0.849	0.6858	0.3191	0.3232	0.5302
33.	či~	80	0.01706	0.736	0.8308	0.3137	0.5873	0.528
34.	foto~	20	0.03106	0.8615	0.6772	0.5128	0.5219	0.5233
35.	poli~	29	0.00165	0.7397	0.8176	0.3053	0.1858	0.5196
36.	trans~	148	0.007972	0.78	0.7639	0.6435	0.6991	0.5173
37.	tene~	16	0.01852	0.9063	0.6147	0.5	0.4771	0.5132
38.	řete~	15	0.001587	0.8626	0.6711	0.3846	0.1633	0.5118
39.	deja~	20	0.06065	0.8517	0.6206	0.3077	0.1728	0.511
40.	media~	17	0.05136	0.8208	0.656	0.5	0.1437	0.5094
41.	ante~	26	0.03611	0.7771	0.7003	0.4194	0.4858	0.5045
42.	tras~	57	0.006543	0.723	0.7817	0.3373	0.28	0.5038
43.	tele~	33	0.004707	0.7202	0.7831	0.4714	0.8265	0.5027
44.	bio~	29	0.008861	0.9022	0.5937	0.3053	0.1541	0.5016
45.	ex~	417	0.01247	0.8421	0.6433	0.3949	0.4459	0.4993
46.	ob~	122	0.005487	0.7562	0.712	0.3333	0.5117	0.4912
47.	řetro~	16	0.002454	0.8491	0.5994	0.4103	0.2683	0.4836
48.	krea~	19	0.02005	0.7444	0.6709	0.4872	0.3019	0.4784
49.	multi~	26	0.001755	0.8323	0.6002	0.4815	0.1934	0.4781
50.	eki~	34	0.002292	0.8385	0.5672	0.3953	0.6393	0.4693

que entre los sufijos, los segmentos con los índices máximos de cuadros sí están incluidos entre los mejores 50. Sin embargo, el ruido no puede ser la única causa de toda la variedad de cosas que aparece en el catálogo de prefijos, especialmente porque también pudo haber afectado al de sufijos, donde los resultados fueron mucho más satisfactorios.

La razón de esta discrepancia está relacionada seguramente con las diferencias entre sufijos y prefijos. Ya se ha observado que en español los sufijos poseen una mayor capacidad gramatical que los prefijos, y que el hecho de que varios de éstos últimos tengan la misma forma que varias preposiciones (*con~*. *en~*, *a~*, etc.) acerca la prefijación a la composición. pero que finalmente hay más semejanzas entre la sufijación y la prefijación que entre esta

última y la composición[48]. Greenberg, por otra parte, caracteriza la diferencia entre sufijos y afijos mediante la tendencia del hablante a anticipar los sonidos[49]. Parece ser que los prefijos tienden a fundirse con la base porque el hablante tiende a anticipar todo aquello que contenga su mensaje, mientras que los sufijos tienden a permanecer relativamente estables porque el hablante matiza con ellos lo ya dicho. Si esta tendencia a anticipar los sonidos es cierta, se explica que los prefijos y las bases se fundan rápidamente —mediante los consiguientes cambios fonológicos—, ocasionando que las bases tiendan a desarrollar muchas más irregularidades que los sufijos, que permanecen relativamente intactos.

Todo esto también está relacionado con el hecho de que, por lo menos en español, tenemos un sistema de sufijos compacto y muy organizado encargado sobre todo de la información sintáctica mientras que los prefijos no cargan información de tipo gramatical, sino de contendido[50]. Así, la entropía al inicio de las palabras (2.55716 bits) es considerablemente más alta que la entropía con que *empiezan* al revés (1.83899 bits).

Por esta razón, el catálogo de prefijos contiene segmentos que no son propiamente verdaderos prefijos del español: además de los pseudo-prefijos tradicionales (*electro~*, *psiko~*, *mikro~*, etc.), hay verbos conjugados en presente de la 3ª p. singular aptos de ocurrir en composiciones (*laba~*, *krea~* y muchos otros distribuidos por todo el catálogo). Cabe señalar que estos últimos no habrían sido aceptados en el catálogo si se hubiera evitado la entrada de segmentos que pudieran tener al interior una frontera de sufijo. De esta manera, *laba~*

[48]Véase Moreno de Alba, *La prefijación en el español mexicano* [106]. UNAM, México, 1996, pp. 15-17.

[49]Greenberg *op. cit.* [58] 1957, pp. 90-93.

[50]Es de sospecharse que en lenguas que no dependan de sus estructuras morfológicas para codificar su información gramatical —o que simplemente no cuenten con un sistema de sufijos— cualquier intento de descubrir sufijos con este método resultaría en un catálogo más parecido al de los prefijos del español, esto es, más voluminoso y poblado de segmentos dudosos y erróneos.

de 'lavamanos' no habría sido seleccionado como prefijo, porque $\sim a$ se puede segmentar de *laba*$\sim$, mientras que *amanos*$\sim$ podría considerarse un sufijo de todo el vocablo, que después de todo no merece aparecer entre los sufijos. porque 'lavamanos' es un ejemplo típico de composición.

Para evitar segmentos flagrantemente erróneos en la tabla 2.13. se filtraron aquellos con un valor de afijalidad menor a 0.45. También se requirió que exhibieran la mejor afijalidad en por lo menos 15 de los vocablos en que ocurren y que fueran el prefijo más exterior en por lo menos 10 vocablos. Además, tienen una probabilidad de ocurrir como prefijos en los vocablos de por lo menos 30% (prob1), y en la cadena de palabras Ψ (el corpus) de al menos 10% (prob2). A todos se les requirió un índice normalizado de de Kock-Bossaert de por lo menos 0.7 y de entropía (también normalizada) de 0.5. Por último, todas las formas ocurren más de 10 veces. De todas maneras y con todas estas restricciones, aparecen formas dudosas que permanecieron por no poderse filtrar automáticamente. Por todo esto y porque el conjunto de sufijos, por el contrario, se mostró mucho más consistente, en la siguiente sección se examinarán con más detalle solamente éstos últimos.

2.8 Los sufijos del español de México

Ya que los sufijos se perfilaron arriba como un subsistema compacto y muy organizado, que, como sabemos, en español se emplea en la codificación de información gramatical, en esta sección se examina el conjunto de cadenas de sufijos presentado arriba. Específicamente, se analizan los segmentos que según este experimento resultaron ser miembros del catálogo

de sufijos. Para esto se examinaron los 749 segmentos más "sufijales" del corpus, que se encuentran consignados en la tabla C.1 del apéndice (a partir de la página 400). En esencia el material de dicha tabla se reorganiza en otras menores para hacer distinciones entre lo que faltó, lo que se consiguió y, especialmente, para indagar la naturaleza de los resultados del procedimiento aplicado en este capítulo. Las tablas presentan grupos sufijales de series de morfos, los menos de los cuales, como se verá, están constituidos de un solo morfo. Por comodidad se organizan en tres grupos: cadenas de sufijos flexivos, de sufijos derivativos y cadenas con enclíticos pronominales.

Sufijos flexivos

Una de las cosas que sobresalen de los sufijos más afijales según la tablas (2.12 y la C.1 del apéndice) es que una sola forma puede representar diferentes morfemas. Por ejemplo, el sufijo *~s* (afjdad. 0.8192) es una marca de plural en sustantivos, pero también una desinencia verbal de 3ª p. singular; y el sufijo *~es* (afjdad. 0.6097), otro alomorfo de la marca del plural en sustantivos, es también flexión verbal de 2ª p. singular. Esa homografía se repite mucho entre las formas que representan morfemas de flexión. Así, aunque la tabla 2.14 contiene sufijos de flexión nominal, es obvio que esos mismos también representan desinencias verbales e, incluso, sufijos derivativos (*~a* y *~o*). Es de destacarse que esta homografía-homonímia no es normalmente un obstáculo para que se distinga el tipo de sufijo en cuestión: el contexto casi siempre es suficiente para desambiguarlo. Tal vez por eso ocurran tantas cadenas de sufijos: se trata de contextos sufijados donde la ambigüedad es menor. Por ejemplo, en *~alidades* no hay ambigüedad en cuanto al carácter que *~es* tiene de ser alomorfo del plural de sustantivo.

Tabla 2.14: Sufijos de flexión nominal

sufijos	fr.[a]	afijalidad
~a	7687	0.8153
~o	6314	0.8222
~s	12013	0.8192
~os	4554	0.7588
~as	4324	0.7547
~es	2479	0.6097

[a]Recuérdese que las frecuencias no se refieren a las de las formas en el corpus, sino al número de vocablos en que obtuvieron el valor de afijalidad más alto (con respecto a las otras segmentaciones posibles de cada vocablo). Evidentemente, como afijos tienen una frecuencia mayor. Y, como formas en el corpus, son todavía más frecuentes.

No es un asunto trivial determinar si la ocurrencia de las formas de la tabla 2.14 al final de un grupo de afijos constituyen la ocurrencia de uno de los morfemas de flexión nominal (masculino, femenino y/o plural) o un pedazo de otro morfema (por ej. en *~amiento* o *~ansia*), pero la prominencia de los sufijos de flexión interviene en la concordancia de sustantivos con artículos, determinadores y adjetivos. aun cuando no ocurren propiamente (sino otros morfemas que casualmente finalizan con *~a* u *~o*).

La cuestión es que aproximadamente 82 de las formas del catálogo terminan en *~o* (con un promedio de afijalidad de 0.5089). Por los contextos, la mayoría son grupos donde sí ocurre el morfema de flexión (*~ado*, *~ero*, *~iko*, *~iyo*, etc.). Similarmente, aproximadamente 77 formas terminan en *~a* (con promedio de afijalidad de 0.50497), donde muchas sí son marcas de género femenino (*~ita*, *~adora*, *~osa*, *~ana*, etc.).

En el caso de la marca de plural nominal hay menos incertidumbre. Hay algo más de 100 secuencias de afijos que contienen el sufijo de flexión nominal de plural. Se trata en su mayoría de algún sufijo derivativo seguido de uno de los alomorfos de. *~s* o *~es* (por ej.. *~ados*, *~antes*, *~adores*, *~itas*, etc.). La afijalidad promedio de esas secuencias es de 0.5211. cantidad nada despreciable, al considerar que más de la mitad de la lista tiene menos.

Pero los sufijos de flexión nominal son pocos. La mejor evidencia de que los segmentos que se aislaron mediante el procedimiento llevado a cabo son en verdad los más afijales es corroborar que el sistema de flexión verbal ocurra lo más completo posible dentro de las 740 formas más afijales de la tabla 2.12 del apéndice. De allí se tomaron los sufijos que aparecen organizados en la tablas 2.15, 2.16 y 2.17. La tabla 2.15 contiene los paradigmas

Tabla 2.15: Sufijos de flexión verbal del modo indicativo

~ar			~er			~ir		
PRESENTE								
~o			6314			0.8222		
~eo	99	0.5342	~es	2479	0.6097			
~as	4324	0.7547						
~eas	15	0.5028						
~a	7687	0.8153	~e	2363	0.6833			
~ea	79	0.5173						
~amos	645	0.6187	~emos	360	0.4888	~imos	151	0.5317
~an	1775	0.6579	~en	945	0.7386			
~ean	25	0.5276						
PRETÉRITO								
~é	639	0.6818	~í	138	0.5279			
~aste	136	0.6051	~iste	70	0.5015			
~ó	1428	0.8428	~ió	303	0.5698			
~eó	21	0.5202						
~amos	645	0.6187	~imos	151	0.5317			
~aron	736	0.6773	~ieron	238	0.5542			
			~eron	11	0.3521			
~ron			317			0.409		

de conjugación verbal del modo indicativo. En la primera parte se exhiben los paradigmas del presente y del pretérito. La segunda contiene los del futuro y la tercera las formas del pospretérito y del copretérito. Esta tabla está dividida en tres columnas, una para cada conjugación. Como en la tabla anterior, cada forma aparece con su frecuencia (como sufijo) y su índice normalizado de afijalidad. Obsérvese que algunas formas no contienen la vocal temática que identifica la conjugación pertinente, por lo que se muestran abarcando las tres columnas. Otras formas son comunes a los paradigmas de la segunda y tercera conjugaciones. por lo que abarcan las dos columnas correspondientes. Nótese que se incluyeron algunas

Tabla 2.15 (continuación):
Sufijos de flexión verbal del modo indicativo

~ar			~er			~ir		
FUTURO								
~aré	127	0.596	[~eré]			~iré	14	0.471
~ré			66				0.437	
~é			639				0.6818	
~arás	46	0.5554	[~erás]			[~irás]		
~rás			32				0.4176	
~ás			36				0.4093	
~ará	387	0.6415	~erá	68	0.4153	~irá	78	0.5177
~rá			462				0.4221	
~á			70				0.3396	
~aremos	104	0.5985	[~eremos]			~iremos	9	0.4426
~remos			62				0.4156	
~emos			366				0.4888	
~arán	256	0.6257	~erán	21	0.396	~irán	52	0.5272
~rán			204				0.4431	
~án			52				0.3749	

formas de paradigmas irregulares, esto es, formas con material adicional (por ej. *~eas* de 'caporaleas', 'bateas', 'tarareas') o más cortas que las regulares (*~eron* de 'fueron', 'trajeron', 'produjeron').

Por otra parte, no se incluyeron formas como *~ua* de 'averigua' (núm. 380 del apéndice[51]), principalmente porque en todo el catálogo no ocurren otros miembros de sus paradigmas. Otras familias de sufijos, sí presentes pero que tampoco se incluyeron en la tabla 2.15 contienen material que no corresponde propiamente ni a la base ni al sufijo (*~go*, *~ga*, *~gas* y *~gan* en verbos como 'oír', 'traer', 'venir', 'tener', etc.; así como *~ka*, *~kas*, *~kan* y *~ska* en los subjuntivos de 'producir' y 'conducir'). Asimismo hay formas de verbos regulares con material adicional, por ejemplo, allí donde una vocal cerrada se adhiere a vocal abierta formando diptongos (*~iar*, *~iaba*, *~iado* e *~iando*).

Aunque también presentes en el catálogo, tampoco se incluyen en estas tablas las marcas

[51] En contraste con *~úa* de 'actúa', 'perpetúa', etc. que no ocurre en el catálogo.

Tabla 2.15 (continuación):
Sufijos de flexión verbal del modo indicativo

~ar			~er			~ir		
POSPRETÉRITO								
~aría	231	0.6198	[~ería]			~iría	43	0.5175
~ría			405		0.4527			
~ía			970		0.5371			
~arías	6	0.4505	~erías	24	0.5118	[~irías]		
~rías			23		0.4593			
~ías			120		0.5244			
~ariamos	36	0.5049	[~eríamos]			[~iríamos]		
~riamos			12		0.3977			
~iamos			96		0.4318			
~arían	81	0.5715	~erían	8	0.5284	~iríán	11	0.4487
~rían			76		0.4183			
~ían			336		0.5226			
COPRETÉRITO								
~aba	828	0.6803	~ía		970		0.5371	
~eaba	13	0.4636						
~abas	30	0.5544	~ías		120		0.5244	
~abamos	115	0.5746	~iamos		96		0.4318	
~aban	551	0.661	~ían		336		0.5226	
~eaban	12	0.4605						

de plural que ocurren al final de varias desinencias verbales y que pueden considerarse morfemas separados; esto es, *~mos*, marca de 1ª p. (afjdad. de 0.4818) y *~n*, marca de 3ª p. y 2ª formal (índice importante de 0.5977).

En cuanto a las flexiones del tiempo futuro, los morfemas correspondientes al verbo 'haber' también aparecen todos en el catálogo, por lo que sí se incluyen en la tabla 2.15 (en la segunda parte). La mayoría tiene una afijalidad comparativamente baja, pero destaca el hecho de que esté presente todo el paradigma (*~é*, *~ás*, *~á*, *~emos* y *~án*).

Algunas formas se muestran entre corchetes cuadrados. Son las que, aunque tienen un lugar en estas tablas, no ocurrieron dentro de las 749 más afijales. Nótese que se trata de formas con uno o más sufijos de flexión adheridos a una vocal temática (por ej. *~eré*, *~irás*, *~eríamos*, etc.). Pero los grupos de sufijos sin las vocales temáticas, que son comunes a las tres conjugaciones, sí están todos (~ré, ~rás, ~remos, ~ría, ~rías, y ~ríamos), cosa que en

cierta medida cubre los huecos.

Tabla 2.16: Flexiones del subjuntivo

~ar			~er/~ir		
PRESENTE					
~e	2363	0.6833	~a	7687	0.8153
~es	2479	0.6097	~as	4324	0.7547
~emos	360	0.4888	~amos	645	0.6187
~en	945	0.7386	~an	1775	0.6579
PRETÉRITO					
~ara	370	0.6467	~iera	165	0.5332
~ra			917	0.5144	
~aras	28	0.5446	~ieras	7	0.5462
~ras			179	0.5157	
~aramos	26	0.48	~ieramos	11	0.4959
~ramos			31	0.4823	
~aran	196	0.6097	~ieran	75	0.5067
~ran			141	0.4873	
~ase	114	0.5925	~iese	9	0.5117
~ases	19	0.5664	[~ieses]		
			~eses	26	0.5613
[~ásemos]			[~iésemos/~ésemos]		
~asen	27	0.5497	[~iesen]		
			~esen	16	0.514
FUTURO					
~are	9	0.4759	~iere	11	0.482
			~ere	12	0.4889
~ares	46	0.4878	~[i]eres	15	0.5178
~res			334	0.4711	
~aremos	104	0.5985	[~iéremos]		
[~aren]			[~ieren]		
~ren			12	0.523	

En la tabla 2.16 están los paradigmas del subjuntivo. De nuevo, las formas faltantes se exhiben entre corchetes cuadrados. Éstas se concentran en el segundo paradigma del pretérito (las formas *~ase* e *~iese*) y en el del futuro (las de *~are* e *~iere*). Ciertamente, los sufijos del presente de subjuntivo están completos en parte porque comparten formas con el presente de indicativo, aunque en diferentes conjugaciones (por ej. los de la primera en indicativo son ahora las de segunda y tercera en subjuntivo). Sin embargo, las del primer paradigma del pretérito, aunque exclusivas a este modo y a este tiempo, están todas. De hecho, la ausencia

de formas del futuro y del segundo paradigma del pretérito es compatible con la intuición de que estén cayendo en desuso.

Tabla 2.17: Sufijos de verboides

sufijos	fr.	afijalidad
~ar	1633	0.6982
~ear	75	0.5405
~er	264	0.5661
~ir	209	0.5938
~r	2587	0.5161
~ando	976	0.6847
~eando	41	0.5189
~iendo	276	0.5738
~ndo	345	0.4111
~ado	1429	0.6917
~eado	23	0.5098
~ido	445	0.6248
~do	2437	0.5092

Por último, los sufijos para formar verboides se muestran en la tabla 2.17. Nótese que los participios pasados se repiten en la tabla 2.18 de la página 151 (en la parte sobre sufijos derivativos) porque también sirven para formar adjetivos a partir de raíces verbales. Cabe notar que en promedio, estos pocos sufijos tienen una alta afijalidad de alrededor de 0.56. Lo interesante es que no falta ninguno e incluso se incluyen formas sin vocales temáticas (*~r*, *~ndo* y *~do*) o con material adicional (*~ear*, *~eando*, *~eado*).

Sufijos derivativos

Como se verá a continuación, los sufijos derivativos (y las cadenas que los contienen) son más numerosos y ocurren por lo general menos a menudo que los de flexión. En ese sentido son menos económicos y, por lo tanto, menos afijales. En esta subsección se organizan en tres grupos, los sufijos derivativos relacionados con el verbo (tabla 2.18), ya sea porque

Tabla 2.18: Los verbos y derivación (con y sin marcas de flexión)

sufijos	fr.	afijalidad
~a	7687	0.8153
~as	4324	0.7547
~o	6314	0.8222
~os	4554	0.7588
~e	2363	0.6833
~es	2479	0.6097
~ado	1429	0.6917
~ao	14	0.5017
~eado	23	0.5098
~ados	941	0.6678
~ada	1135	0.6768
~eada	21	0.5247
~adas	813	0.6602
~ido	445	0.6248
~idos	269	0.603
~ida	304	0.6022
~idas	218	0.581
~da	441	0.4732
~das	174	0.4713
~do	2437	0.5092
~dos	214	0.4616
~ando	976	0.6847
~*isó*	15	0.4805
~*isar*	69	0.4903
~*isarse*	7	0.3989
~*isando*	9	0.4577
~*isada*	23	0.4651
~*isadas*	12	0.5005
~*isado*	35	0.4861
~*isados*	17	0.4521
~*ifika*	12	0.4155
~*ifikada*	7	0.462

convierten raíces verbales en sustantivos o adjetivos. o porque convierten raíces no verbales (por ej. adjetivos) en verbos; grupos de sufijos adverbiales (tabla 2.19); y sufijos derivativos nominales (el resto de las tablas).

En cuanto a la derivación que involucra verbos como bases o que resulta en la formación de verbos, en la tabla 2.18 se exhiben, primero, algunos sufijos que se adhieren a verbos para formar sustantivos, 'compra', 'logro', 'corte'; luego, marcas participiales para formar adjetivos y sustantivos a partir de bases verbales; y, finalmente, sufijos que se adhieren a sustantivos o

adjetivos para formar verbos ('mit**ificar**', 'escen**ificar**'...; 'neutral**izar**', 'plural**izar**', etc.).

Algunas formas, *~itar*, *~itan* y el participio *~itado* obtuvieron un grado de afijalidad cercano a 0.5 (0.493, 05214 y 0.5099 respectivamente), lo que invitaba a considerar la posibilidad de incluirlas en esta tabla. Sin embargo, al examinar el tipo de vocablos de donde provienen ('editar', 'dinamitar', 'gravitar', 'evitar', 'meditar', 'necesitar', etc.), sus reducidas frecuencias de ocurrir en éstos como afijos (7, 6 y 8 respectivamente), y no encontrar algún significado que lo justifique como morfema, se excluyeron de dicha tabla.

Con respecto a la derivación de adverbios, en la tabla 2.19 se reúnen las secuencias de sufijos que contienen *~mente*. Lo que salta a la vista es el tipo de sufijos a los que se adhiere. Como era de esperarse, todos son sufijos derivativos para formar adjetivos.

Tabla 2.19: Grupos de sufijos con marca adverbial

sufijos	fr.	afijalidad
~mente	981	0.4728
~ablemente	6	0.4996
~amente	624	0.6169
~adamente	53	0.5457
~almente	74	0.5171
~atibamente	8	0.5004
~tibamente	10	0.4634
~emente	7	0.3911
~idamente	9	0.4801
~ikamente	63	0.5123
~osamente	34	0.516

En cuanto a la derivación nominal, el conjunto de secuencias de sufijos derivativos es enorme, cosa que se refleja en la tabla C.1 del apéndice. Como estrategia para revisarlos de manera breve pero ordenada, los cotejaré con uno de los catálogos compilados por Moreno de Alba, concretamente, aquel de sufijos ordenados por su forma[52], que no consigna todos los

[52]Capítulo V del libro *Morfología derivativa nominal en el español de México*, [105], pp. 183-205.

sufijos analizados por ese investigador, pero sí algunos de los más importantes agrupados por su semejanza formal y ordenados por porcentaje de ocurrencias en su material. La especial pertinencia de la semejanza formal es obvia, ya que el conjunto de sufijos extraídos del *CEMC* es, en esencia, un conjunto de formas (cadenas de morfos). De esta manera, en la tabla 2.20 se agrupan por parecido formal varias secuencias de sufijos que contienen por lo menos uno de derivación nominal. Estos sufijos de derivación, además de ocurrir en las secuencias de morfos determinadas en este experimento, se documentan en alguno de los grupos de la lista de Moreno de Alba.

Obviamente, aquí no hay espacio para examinar toda la morfología derivativa del español de México. De todas maneras, los demás sufijos derivativos que se descubrieron en este experimento aparecen sin comentario alguno en la tabla C.2 del apéndice (a partir de la página 419, después de la de las 749 cadenas de sufijos más afijales). Varios de esos sufijos son más prominentes que los que a continuación se examinan. De hecho, las secuencias de sufijos agrupadas por significado muestran obviamente mucha mayor homogeneidad de sentido. Por eso, los de la tabla C.2 (de sufijos descubiertos que se dejaron en el apéndice), que están agrupados improvisadamente por formas y significados, contienen sufijos derivativos tanto o más prominentes (por ej. *~ismo*, *~oide*, *~ísimo*, *~eño*, etc.) que las que se examinan a continuación.

Como se dijo, Moreno organiza sus sufijos por semejanza formal. En ese sentido los grupos son alomorfos del morfema en cuestión, aunque, como dice el investigador, a veces sea difícil argumentar que conserven una "aceptable homogeneidad de sentido"[53]. Una diferencia

[53] *Ibid.* [105], p. 183.

Tabla 2.20: Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	fr.	afijalidad
~(V)(C)ión	~asión	540	0.6007
	~*ali*sasión	18	0.4903
	~isión	39	0.4782
	~tasión	23	0.4727
	~lasión	6	0.4585
	~isasión	106	0.4485
	~*ifik*asión	23	0.4395
	~sión	863	0.4213
	~csión	6	0.4042
	~ión	41	0.2938
~V	~a	7687	0.8153
	~e	2363	0.6833
	~o	6314	0.8222
	~as	4324	0.7547
	~os	4554	0.7588
	~es	2479	0.6097
	~ea	79	0.5173
	~eo	99	0.5342
	~eos	18	0.5192

importante entre los sufijos de Moreno y los de la tabla 2.20 es que éstos últimos representan secuencias de sufijos, entre los que se encuentra el alomorfo en cuestión (Moreno los presenta aislados). Evidentemente, a partir de estos grupos se puede investigar de manera automática la afitáctica de todas y cada una de las secuencias y, por lo tanto, del español de México (tanto de morfos de flexión como de derivación). Pero por ahora, dada la complejidad de esa tarea, esto queda pendiente.

El primer grupo[54], ~(V)(C)ión (*~ación, ~ión*), que sirve para formar sustantivos de acción o efecto, fue el más frecuente de su material (12% de su total de vocablos). Esto, al considerar las frecuencias, no coincide con los datos obtenidos en este trabajo (compárense las de este grupo con las del segundo).

De hecho, el orden tampoco corresponde al de afijalidad, según se ha definido aquí. Por

[54]En la notación de Moreno de Alba, 'V' significa cualquier vocal, 'C' cualquier consonante y '-' flexión nominal de género. Los paréntesis indican que un elemento puede no estar presente.

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	fr.	afijalidad
~(V)(C)it-	~ita	453	0.6219
	~ito	421	0.6093
	~*a*dito	10	0.5316
	~*a*dita	21	0.5303
	~*e*sito	20	0.5165
	~*e*sita	10	0.5051
	~*on*sito	14	0.5001
	~sita	18	0.4834
	~sito	64	0.4629
	~rito	9	0.4561
	~tito	8	0.4528
	~itos	271	0.5984
	~itas	210	0.5962
	~*a*ditas	7	0.5045
	~*e*sitos	7	0.4658
	~sitos	28	0.476
~(V)al	~al	375	0.5515
	~ual	12	0.4362
	~ial	32	0.4268
	~ales	281	0.5509
	~uales	9	0.4449
	~iales	10	0.4756

ejemplo, el promedio de afijalidad del primer grupo es bajo (0.45077) al compararlo con el del grupo siguiente, ~V (vocales para formar sustantivos con sentido de acción o efecto: ~*e*, ~*o*, y con material adicional, ~*eo*), que es 0.7736 (tomando en cuenta solamente los sufijos de una sola vocal y sin marca de plural). Es más, el grupo de alomorfos de diminutivo, ~(V)(C)it- (~*adito* ~*itito*, ~*ita*), que sigue a los dos primeros (en la segunda parte de la tabla, una página después) muestra frecuencias menores y tiene un respetable promedio de afijalidad de 0.5155.

Nótese que el segundo grupo comparte formas con el paradigma de flexión nominal de género, lo que explica el alto promedio de afijalidad, pero hay varios grupos con promedios de afijalidad más alta que el primero. En suma, el orden de importancia de Moreno no concuerda ni con los datos de frecuencia ni con la medida de afijalidad calculadas en este trabajo. Algunos ejemplos de vocablos que exhiben los sufijos de los primeros tres grupos son:

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	fr.	afijalidad
~(V)(C)(C)ic-	~ika	288	0.5585
	~iko	341	0.5581
	~ikas	167	0.5551
	~ikos	193	0.5544
	~tika	46	0.5028
	~*ist*ika	9	0.5024
	~*on*iko	11	0.5015
	~*on*ika	11	0.5001
	~tiko	45	0.4989
	~riko	7	0.4922
	~*ist*iko	12	0.4691
	~niko	17	0.4429
	~*if*ika	12	0.4155
	~tikas	27	0.5104
	~tikos	34	0.48
~(V)(C)(C)ad	~idad	334	0.5214
	~*os*idad	11	0.5154
	~*ad*idades	8	0.5035
	~idades	86	0.4865
	~*ad*idad	50	0.4771
	~*abil*idad	19	0.4447
	~sidad	28	0.4268
	~edad	24	0.4211
	~*is*idad	11	0.3913
	~ridad	9	0.3861
	~nidad	7	0.3776
	~lidad	8	0.309
	~*bil*idad	7	0.2957
	~dad	18	0.2826
	~ad	22	0.2517
	~*il*idad	8	0.2316

'aclar**ación**', 'nacion**alización**', 'admis**ión**'; 'siembra', 'enfoque', 'muestr**eo**'; y 'cafe**cito**', 'sent**adito**', 'cabez**oncito**', etc.

El grupo siguiente, representado por ~(V)al (*~al, ~ual*), sirve típicamente para formar adjetivos que designan relación o caracterización. Algunos ejemplos son 'experiment**al**', 'mund**ial**' y 'man**ual**'.

Luego, los sufijos del grupo ~(V)(C)(C)ic- (*~ico, ~ística*) forman adjetivos y sustantivos con sentido técnico; por ejemplo, 'magn**ífico**', 'electr**ónica**' y 'caracter**ístico**'.

El conjunto de sufijos con el esquema ~(V)(C)(C)ad (*~idad, ~dad*) suelen formar sustantivos abstractos. Algunos ejemplos son: 'carn**osidad**', 'nacion**alidad**', 'suav**idad**', 'crue**ldad**' y 'puber**tad**'.

El grupo siguiente, representado por ~(V)Vnte (*~ante, ~iente*), sirve para formar adjetivos a partir de verbos; por ejemplo, 'ayud**ante**', 'conven**iente**' y 'absorb**ente**'.

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**fr.**	**afijalidad**
~(V)Vnte	~ante	240	0.5998
	~antes	187	0.6037
	~entes	70	0.5146
	~ientes	26	0.5048
	~iente	51	0.4821
	~ente	84	0.3721
	~nte	212	0.3659
	~ntes	40	0.3591

El siguiente conjunto, ~Vd- (*~ado, ~ida*), corresponde al grupo presentado arriba en la tabla 2.18. Se trata de varias secuencias de sufijos, uno de los cuales es una forma participial con, por lo menos, vocal temática y marcas de flexión nominal. Algunos ejemplos de vocablos formados por este grupo son: 'señ**alado**', 'part**ida**', 'subordin**ado**' y 'clas**ificada**'.

El siguiente grupo de la tabla 2.20 es el representado mediante ~Vncia o ~anza (*~ancia, ~anza*) se utiliza para formar sustantivos de acción o resultado de la acción; por ejemplo, 'mat**anza**', 'toler**ancia**' y 'apar**iencia**'.

Luego está el conjunto representado por ~(u)os- (*~oso, ~uosa*). Este grupo sirve para formar adjetivos de cualidad o defecto. Algunos ejemplos son: 'fam**oso**', 'mald**oso**', 'defect**uoso**' y 'juic**ioso**'.

El grupo ~(Vd)er- (*~adero, ~era*) sirve para formar sustantivos y adjetivos que desig-

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	fr.	afijalidad
~Vd-	~ado	1429	0.6917
	~ada	1135	0.6768
	~ados	941	0.6678
	~adas	813	0.6602
	~ido	445	0.6248
	~idos	269	0.603
	~ida	304	0.6022
	~idas	218	0.581
	~*al*ado	6	0.5669
	~*or*ado	6	0.5549
	~iados	16	0.5289
	~eada	21	0.5247
	~iada	14	0.5193
	~*in*ado	7	0.5146
	~esido	10	0.5099
	~eado	23	0.5098
	~*is*adas	12	0.5005
	~iado	31	0.5004
	~*is*ado	35	0.4861
	~*en*ado	6	0.4827
	~*on*ada	6	0.4806
	~iadas	10	0.4736
	~*on*ado	8	0.4714
	~*on*adas	8	0.4673
	~*is*ada	23	0.4651
	~*ifik*ada	7	0.462
	~*is*ados	17	0.4521

nan algún agente, instrumento, objeto, alimento, etc. Por ejemplo, 'sal**ero**', 'limon**ero**' y 'pan**adera**'.

El conjunto con la forma ~Vm(i)ent- (~*amento*, ~*imienta*) forma sustantivos de acción o resultado. Por ejemplo, las voces 'carg**amento**', 'herr**amienta**' y 'mov**imiento**'.

El siguiente grupo está representado por ~(Vt)iv- (~*ativo*, ~*iva*). Sus formas sirven para formar adjetivos. Algunos ejemplos son: 'signific**ativa**', 'conflict**ivo**', 'consec**utivo**' y 'expres**ivo**'.

Luego está el conjunto de formas representado por ~Vble (~*able*, ~*ible*) y que para formar adjetivos que suponen capacidad y aptitud. Por ejemplo, 'inaplaz**able**', 'sens**ible**' y

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	fr.	afijalidad
~Vncia ~anza	~ansas	14	0.5525
	~ansa	23	0.5138
	~ensias	19	0.5059
	~ansia	22	0.4916
	~ensia	83	0.4761
	~iensia	6	0.4024
	~nsia	32	0.3529
~(u)os-	~osa	178	0.5802
	~oso	191	0.5732
	~osos	112	0.5673
	~osas	91	0.5653
	~ioso	8	0.4332
~(Vd)er-	~ero	292	0.5953
	~era	335	0.5569
	~eros	200	0.5867
	~eras	137	0.5822
	~*on*ero	10	0.5168
	~*on*eros	8	0.5224
	~*ad*ero	16	0.5439
	~*ad*eros	6	0.5216
	~*ad*eras	6	0.4727
	~dero	11	0.3831

'insolu**ble**'.

El grupo ~(V)Cor- (notación que debería tener la marca de flexión de género entre paréntesis ~(V)Cor(-): *~ador*, *~idor*, *~tor*) sirve para formar adjetivos que designan oficios, profesiones y ocupaciones. Algunos adjetivos son: 'pirograb**ador**', 'profer**idora**' y 'protec**tora**'.

El conjunto de formas representado mediante ~a(ta)ri- (*~aria*, *~atario*) se utiliza para formar sustantivos y adjetivos con significados típicamente colectivos, locativos, etc. Por ejemplo, 'origin**ario**', 'universi**taria**' y 'prole**tario**'.

El grupo de la notación ~í- (*~ío*, *~ía*) se utiliza para formar sustantivos abstractos. Ejemplos de estos sustantivos son: 'mejor**ía**', 'arque**ología**', 'burgue**sía**' y 'jud**ío**'.

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**fr.**	**afijalidad**
~V'm(i)ent-	~amiento	141	0.6005
	~amientos	47	0.5506
	~imientos	9	0.5103
	~imiento	83	0.4751
	~amento	15	0.4292
	~miento	248	0.35
	~mento	7	0.3412
	~mientos	12	0.3068
~(Vt)iv-	~atibo	55	0.5418
	~atiba	44	0.5382
	~atibas	32	0.5286
	~atibos	30	0.5219
	~tibo	63	0.4868
	~tiba	54	0.4888
	~tibos	37	0.5066
	~tibas	39	0.4866
	~iba	15	0.422
~Vble	~able	115	0.5679
	~ables	88	0.5611
	~ibles	13	0.4713
	~ible	31	0.4652
	~ble	87	0.3569
	~bles	16	0.2973

El conjunto siguiente, ~(V)(C)ón(-) (*~ón, ~ona*), se emplea para construir sustantivos y adjetivos aumentativos o de acción contundente. Algunos ejemplos son: 'apret**ón**', 'pis**otón**', 'saca**tones**' y 'empuj**oncito**'.

El grupo representado mediante ~Vría (*~aría, ~ería*) se utiliza en la formación de sustantivos. Por ejemplo, sustantivos como 'secret**aría**', 'not**aría**', 'enferm**ería**' y 'ingenier**ía**'.

Luego está en conjunto ~(V)(C)ura (*~ura, ~adura*) para formar sustantivos. Algunos sustantivos formados con estos sufijos son: 'pint**ura**', 'criat**ura**' y 'colg**adura**'.

El conjunto ~(V)(C)ez(a) (*~ez, ~aleza*) sirve para formar sustantivos a partir de adjetivos. Por ejemplo, los sustantivos 'madur**ez**', 'trist**eza**' y 'estupid**eces**'.

El grupo siguiente es el representado por ~(V)(C)ori- (*~orio, ~atoria*). Estas formas

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	**fr.**	**afijalidad**
~(V)Cor(-)	~ador	268	0.6147
	~edor	12	0.4883
	~idor	12	0.5487
	~adora	124	0.581
	~adores	196	0.6033
	~idores	11	0.4858
	~adoras	41	0.5716
	~tores	47	0.5069
	~toras	10	0.4947
	~tora	7	0.4837
	~tor	51	0.4832
	~dor	108	0.3902
	~dores	77	0.3736
	~dora	32	0.3407
	~or	324	0.4573
	~ora	146	0.4793
	~ores	189	0.4584
	~oras	26	0.4649
~a(ta)ri-	~aria	44	0.5133
	~arios	51	0.5059
	~arias	17	0.5017
	~ario	76	0.4935
	~tario	6	0.4214
~í-	~ía	970	0.5371
	~*esía*	8	0.5283
	~ías	120	0.5244
	~sía	12	0.4639
	~íos	7	0.4604
	~*olojía*	18	0.445
	~ío	16	0.4232

se utilizan para formar sustantivos y adjetivos. Algunos ejemplos son: 'reclin**atorio**', 'escap**atoria**' y 'dorm**itorio**'.

Enseguida está el conjunto ~Vdor(a) (*~ador*, *~edor*, *~idora*) para formar sustantivos que designan objetos, instrumentos o lugares. Por ejemplo, los sustantivos 'incub**adora**', 'corr**edor**' y 'med**idor**'.

El conjunto de formas representado por ~in- (*~ino*, *~ina*) sirve para formar sustantivos muy diversos y adjetivos que señalan semejanzas y características. Algunos ejemplos son: 'alcal**ino**', 'cervant**ino**' y 'estudiant**inas**'.

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	fr.	afijalidad
~(V)(C)ón(-)	~ón	957	0.443
	~ones	815	0.4777
	~ona	95	0.5453
	~onas	20	0.5269
	~tón	12	0.4346
	~tones	10	0.4397
	~on*eros*	8	0.5224
	~on*ero*	10	0.5168
	~on*sito*	14	0.5001
~Vría	~aría	231	0.6198
	~ería	121	0.5364
	~erías	24	0.5118
	~rías	23	0.4593
	~arías	6	0.4505
~(V)(C)ura	~ura	88	0.4577
	~uras	24	0.4632
	~adura	11	0.4739
	~aduras	11	0.5024
	~tura	19	0.4531
	~turas	7	0.4987
	~atura	9	0.4414

El siguiente grupo es muy variado y está representado por la notación ~t- (*~te*, *~to*). Estas formas dan lugar a sustantivos y adjetivos abstractos de acción o efecto. Algunos vocablos con estos sufijos son: 'muer**te**', 'aten**to**' y 'ven**ta**', 'produ**cto**' e 'institu**to**'.

El conjunto representado mediante ~(i)(t)ud (*~itud*, *~ud*) sirve para formar sustantivos también abstractos que designan cualidad, acción o conducta. Por ejemplo, los sustantivos 'exact**itud**', 'ineptitud' y 'juventud'.

El antepenúltimo conjunto es el representado por ~(c)ill- (*~illo*, *~illa*). Este grupo se utiliza típicamente para formar sustantivos y adjetivos despectivos y a veces diminutivos. Algunos ejemplos son: 'chiqu**illo**', 'cortin**illa**' y 'molin**illo**'.

El penúltimo grupo está constituido por ~(i)ci- (*~icio*, *~cia*). Éstas formas se unen a raíces adjetivas para formar sustantivos abstractos. Por ejemplo, las voces 'inmundi**cia**'.

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	fr.	afijalidad
~(V)(C)ez(a)	~és	89	0.5357
	~esa	51	0.5368
	~eses	26	0.5613
	~esas	24	0.5551
~(V)(C)ori-	~atoria	22	0.5203
	~atorios	8	0.5195
	~atorias	8	0.4597
	~atorio	13	0.4477
	~torio	10	0.4172
~Vdor(a)	~ador	268	0.6147
	~edor	12	0.4883
	~idor	12	0.5487
	~adores	196	0.6033
	~idores	11	0.4858
	~adora	124	0.581
	~adoras	41	0.5716
	~dor	108	0.3902
	~dores	77	0.3736
	~dora	32	0.3407
~in-	~ino	48	0.55
	~inas	38	0.5498
	~ina	115	0.5483
	~inos	34	0.5356

'silen**cio**' e 'infan**cia**'.

El último de los grupos organizados por forma es ~i- (*~ia. ~ie. ~io*) y es el menos documentado entre los materiales de Moreno de Alba (0.4% de su total). Las formas de este grupo dan lugar a sustantivos abstractos relacionados con verbos o nombres. Algunos ejemplos son los sustantivos 'molest**ia**', 'domin**io**' y 'proge**nie**'.

Como se puede corroborar al examinar el catálogo de sufijos del apéndice, las formas que Moreno de Alba consignó en su estudio no agotan todo sufijo o grupo de sufijos derivativos que se consignan en el catálogo del apéndice. Hasta aquí se ha mostrado que los fragmentos de la tabla del apéndice no son cualquier cosa. De hecho, todavía quedan allí muchas que también se pueden agrupar por forma. Algunos otros grupos se presentan después de la tabla del apéndice a partir de la página 419 en la tabla C.2.

Tabla 2.20 (continuación):
Grupos de sufijos derivativos nominales (según parecido formal)

	sufijos	fr.	afijalidad
~t-	~to	366	0.4879
	~ta	604	0.5061
	~te	1072	0.4769
	~tas	254	0.5062
	~tos	176	0.5062
	~tes	183	0.4582
	~rte	83	0.4315
	~ste	118	0.3839
	~cto	6	0.3832
	~sta	11	0.3585
	~nta	7	0.3336
	~uto	9	0.541
	~ato	60	0.5369
~(i)(t)ud	~itud	14	0.3739
~(c)ill-	~iya	107	0.524
	~iyo	75	0.5157
	~iyas	54	0.5258
	~iyos	36	0.5144
~(i)ci-	~sio	11	0.4271
	~sia	154	0.365
	~isia	9	0.4544
~i-	~ia	204	0.4948
	~io	182	0.478
	~ie	13	0.4902
	~ias	60	0.4842
	~ios	64	0.4865

Por otra parte, si comparamos el promedio de cantidades normalizadas de afijalidad de los sufijos derivativos (tablas 2.18 y 2.20) con aquellos que ocurren en el verbo (tablas 2.15, 2.16 y 2.17), encontramos que los primeros (afjdad. 0.48641572) son menos afijales que los segundos (afjdad. 0.5350472). Hay muchas razones por qué estos promedios son más bien burdos, pero a grandes rasgos van de acuerdo con la intuición de que los verbales, los más frecuentes pero con menos tipos o formas, son los más afijales. Los derivativos son un conjunto de más formas (más de 200) con menos ocurrencias cada una. De hecho, mientras en la tabla 2.20 tenemos una selección de sufijos derivativos larga e incompleta, los tipos de segmentos de flexión que faltan no sólo son pocos, sino que además sabemos exactamente cuáles faltan (como ya se dijo, están marcados en las tablas de arriba mediante corchetes cuadrados).

Enclíticos

Como se mencionó arriba, al adherirse gráficamente a la palabra escrita, los enclíticos tradicionales ocurren necesariamente entre los sufijos gráficos. Al considerar las semejanzas entre clíticos y afijos, esto no debe ser una situación incómoda[55]. La tabla 2.21 muestra aquellos segmentos que consisten únicamente de enclíticos y que aparecen dentro de las 600 formas más afijales. Nótese que, de todas las combinaciones de enclíticos posibles, solamente

Tabla 2.21: Enclíticos descubiertos como sufijos gráficos

encls.	fr.	afijalidad
~me	565	0.4744
~te	1072	0.4769
~se	1619	0.5332
~la	633	0.5306
~las	262	0.5171
~le	613	0.526
~les	455	0.4973
~lo	792	0.5384
~los	410	0.5182
~nos	361	0.4643
~selo	7	0.4472

~selo (núm. 598) ocurre en la tabla del apéndice. Lo importante es que por lo menos todos los enclíticos están aquí (la forma correspondiente al pronombre *~os* del dialecto peninsular, no ocurre en este corpus como tal, sino como marca de nombre masculino y plural). Es interesante que después del acusativo *~lo*, sea *~se* el más afijal.

Pero la mayoría de las ocurrencias de enclíticos en la tabla ocurrió adherida a sufijos que se utilizan para marcar gerundios, imperativos e infinitivos. La tabla 2.22 muestra los sufijos

[55] De hecho, no es extraño que ya se hayan analizado los pronombres clíticos del español como marcas de flexión del verbo en concordancia con los complementos; véase el tercer capítulo de Rini, *Motives for Linguistic Change in the Formation of the Spanish Object Pronouns* [119], Juan de la Cuesta, Newark, Delaware, 1992. Por otra parte, véase también la discusión sobre los pronombres personales del portugués como formas de flexión en Spencer, *op. cit.* [129] 1991, p. 382: "if the pronoun forms really are inflections, it's hard to see what grammatical function they have".

del gerundio que ocurrieron con algún enclítico. El único paradigma completo es el de las

Tabla 2.22: Gerundio y enclíticos

gerundio + encl.	fr.	afijalidad
~andome	48	0.5512
~andote	14	0.4973
~andose	260	0.6262
~andole	75	0.5417
~andoles	13	0.4637
~andolo	78	0.564
~andolos	44	0.5324
~andola	58	0.5584
~andolas	22	0.5356
~andonos	18	0.4904
~iendose	68	0.485
~iendolo	12	0.4709
~ndole	14	0.3988
~ndose	43	0.3986

formas con vocal temática de la primera conjugación. Destaca el gerundio en ~se por ser el más afijal y más frecuente como sufijo. De la segunda y tercera solamente hay dos segmentos, a los que se adhieren los enclíticos *~se* y *~lo*. Las formas sin vocal temática también son solamente dos. Es significativo que los segmentos de los paradigmas incompletos tengan las afijalidades más bajas. También es de notarse que en general todas tengan frecuencias bajas como sufijos.

Tabla 2.23: Imperativo y enclíticos

~ar			~er/~ir		
~ame	74	0.5687	~eme	21	0.5052
~ate	107	0.5637	~ete	59	0.5398
~ese	143	0.5256	~ase	114	0.5925
~ala	30	0.5727	~ela	48	0.5296
~alas	14	0.5817	~elas	20	0.4819
~ale	48	0.549	~ele	12	0.5112
~ales	281	0.5509	~eles	11	0.4673
~alo	45	0.5639	~elo	60	0.5194
~alos	32	0.5458	~elos	31	0.5301
~anos	53	0.5396	~enos	6	0.521
plural					
~ense	30	0.5079	~anse	11	0.5291
~nse		7		0.3949	

En la tabla 2.23 se listan los segmentos con sufijos de imperativo y enclíticos. Los paradigmas están completos (sólo aquellos de un enclítico). Es de notarse, sin embargo, que algunas formas son homónimas de otros sufijos, especialmente derivativos (por ej., la secuencia *~anos* de 'campiranos').

Finalmente, la tabla 2.24 contiene las ocurrencias de infinitivo junto a enclíticos. Como se ve, los paradigmas de la segunda y tercera conjugaciones no están completos. Curiosamente

Tabla 2.24: Infinitivo y enclíticos

	~ar		~er			~ir			~r		
~arme	244	0.6306	~erme	48	0.4247	~irme	23	0.5322	~rme	202	0.4144
~arte	144	0.6045	~erte	18	0.466	~irte	8	0.5465	~rte	83	0.4315
~arse	665	0.6692	~erse	105	0.502	~irse	108	0.5659	~rse	786	0.4243
~arla	270	0.635	~erla	22	0.5174	~irla	23	0.5209	~rla	93	0.3926
~arlas	139	0.6021	~erlas	6	0.4816	~irlas	8	0.524	~rlas	18	0.3998
~arle	176	0.5997		[~erle]		~irle	12	0.481	~rle	56	0.4228
~arles	72	0.5505		[~erles]			[~irles]		~rles	14	0.3824
~arlo	316	0.6356	~erlo	35	0.4729	~irlo	39	0.5127	~rlo	140	0.4128
~arlos	201	0.6153	~erlos	15	0.4823	~irlos	16	0.518	~rlos	42	0.3711
~arnos	139	0.5917	~ernos	11	0.4868	~irnos	18	0.4576	~rnos	115	0.3932
~isar*se*	7	0.3989									
~arselo	10	0.4553									

las formas que faltan son del dativo. Si bien, la presencia del paradigma de las formas sin vocales temáticas cubre el espacio que dejan las formas faltantes, es de notarse que en todos los paradigmas tienen valores bajos de afijalidad, especialmente aquellos con enclítico *~les*. De hecho, en las tablas anteriores se observa un poco de lo mismo, aunque mucho menos pronunciado. Por último, destaca la ausencia casi total de combinaciones de enclíticos. Sólo está la secuencia *~selo* en el grupo de afijos *~arselo* (la misma combinación de enclíticos que se exhibe en la tabla 2.21).

El promedio de afijalidad de los 87 segmentos que contienen algún enclítico es de 0.51, cosa significativa al observar que las formas de alrededor de la mitad de los segmentos en la lista

del apéndice tienen una afijalidad menor. Esto es curioso porque los sufijos de flexión tienen como promedio 0.53 y los derivativos 0.49. Entonces, según los criterios cuantitativos de afijalidad propuestos en este capítulo, los enclíticos son menos afijales que los sufijos flexivos pero más que los derivativos. Esto no es decir que sean algo intermedio. Conceptualmente, la enclisis de pronombres se aleja mucho del fenómeno de derivación. Pero que se trate de un grupo finito de pronombres con un número finito de combinaciones posibles (aunque esas combinaciones aquí no se hayan manifestado) y que, incluso, en análisis previos ya han sido considerados marcas de flexión (aunque no exhiban otra función gramatical que la de servir de anáforas), son motivos para no extrañarse por el promedio más bien alto de afijalidad que obtuvieron.

Hacia el catálogo de sufijos del español de México

Como se ve, hasta aquí se logró reunir la gran mayoría de los sufijos de flexión y un conjunto significativo de los derivativos[56]. A partir de esto se puede construir una caracterización completa y sistemática del conjunto de sufijos del español de México, lo que no era el objetivo de este trabajo. Es decir, la tarea exhaustiva de examinar cada uno para determinar su comportamiento y su significado en cada contexto —mediante, por ejemplo,

[56]El procedimiento también se aplicó a un corpus minúsculo (archivo plano de 86 KB) con 15,485 palabras gráficas (poco más de de 2,300 vocablos) de la lengua chuj que Cristina Buenrostro (Corpus de la lengua chuj [23], archivo plano, 2002) amablemente tuvo a bien proporcionarme. El objetivo de ese experimento fue determinar si, con tan pocos datos, por lo menos la morfología flexiva del chuj era susceptible de descubrirse. Los resultados fueron muy alentadores (se presentaron en la ponencia "Características cuantitativas de la morfología flexiva del chuj" [99] en el VII Encuentro Internacional de Lingüística en el Noroeste, noviembre, 2002). Alrededor del 86% de los sufijos y prefijos de flexión (3 prefijos de tiempo, 4 pronombres absolutivos prefijados, 11 ergativos también prefijados y 7 sufijos de voz pasiva y antipasiva, uno de modo, otro de tiempo y dos vocales que marcan final de frase) ocurrieron entre los más afijales de ese pequeño corpus. Es más, de entre los 200 fragmentos examinados, todos los afijos de flexión identificados ocurrieron apretados en los primeros lugares de los dos catálogos construidos (uno de prefijos y otro de sufijos).

concordancias— queda fuera de este trabajo.

También resalta la necesidad de investigar cuantitativamente el fenómeno de parasíntesis. es decir, la co-ocurrencia de ciertos prefijos y ciertos sufijos para formar vocablos nuevos[57]. Esto depende naturalmente de la investigación previa y completa del catálogo de prefijos cuantitativamente reconocibles a partir del corpus.

Además, los sufijos derivativos descubiertos en el *CEMC* mediante el procedimiento aplicado en este capítulo merecen examinarse todavía con más profundidad, lo que también queda fuera de esta tesis. A pesar de que la mayoría de las formas de la tabla C.1 del apéndice se pueden reorganizar para mostrar su pertinencia dentro un subsistema morfológico flexivo y uno léxico derivativo, es necesario analizar lo que aparentemente no entra en ningún lado. Es decir, aunque se observan varios segmentos que se reconocen como elementos con significado que se adhieren a otros para formar nuevos elementos, hay muchos que sólo con un análisis más detallado, específicamente, examinando los vocablos y contextos donde ocurren, podrán considerarse verdaderos morfemas del español.

2.9 Observaciones finales

En este último apartado se resumen las ideas presentadas y los logros y problemas de su aplicación al *CEMC*. En este capítulo, se describió un camino posible para la construcción de dos catálogos de afijos para lenguas similares al español. Y si los miembros de este tipo de catálogos son afijos, no es porque alguien en particular haya decidido que deberían ser

[57] Véase caracterización de construcciones parasintéticas, por ejemplo en Moreno de Alba, *op. cit.* [106] 1996, pp. 31-37.

tratados como tales, sino porque exhiben ciertas propiedades lingüístico-formales que hemos asociado al concepto de afijo. En concreto, los afijos son muy pocos pero muy frecuentes y se adhieren a numerosos segmentos de baja frecuencia (bases) para formar muchos signos del nivel léxico (satisfaciéndose así la necesidad de un inventario económico de signos). Así, estos signos léxicos se relacionan los unos con los otros principalmente porque comparten afijos. De hecho, un afijo se combina con muchas bases que también se combinan con otros pocos afijos para formar todavía más signos léxicos (cada afijo participa en un número determinado de *cuadros*). Finalmente, la ocurrencia de un afijo debe ocasionar menos sorpresa (en unidades de entropía), que la ocurrencia de una base.

Las ideas presentadas aquí dan todavía mucho lugar para perfeccionarse y muchas cuestiones merecen ser examinadas con más detenimiento. Por un lado, está el asunto del mejoramiento de las técnicas de programación (por ej., cómo acelerar las cuentas de cuadros). Por el otro, hay muchas otras cuestiones relacionadas con la manera de acercarse al fenómeno (asuntos de interés lingüístico). Por ejemplo, cómo se deben contar los cuadros y si deberían considerarse otras estructuras combinatorias (tales como XAZ, XBZ, XCZ, etc.). Además, también la medida de economía debe investigarse con más detenimiento (¿qué relación tiene esta medida con el número de cuadros en la segmentación analizada?). Asimismo, hay mucho que explorar con respecto al fenómeno de la información (¿cuál es la mejor manera de medir la entropía dentro de la palabra?, ¿qué se debe considerar para su cálculo, las frecuencias absolutas en el corpus o solamente el número de formas?, ¿hay alguna manera de combinar las entropías en ambas direcciones para determinar con más seguridad fronteras morfológicas?). Finalmente, ¿hay alguna mejor manera de combinar estos índices, o existen otros mejores

que se puedan combinar para caracterizar cuantitativamente la propiedad morfológica de afijalidad?, ¿cómo se deben tratar los errores?, o más bien, ¿qué es propiamente un error?

Por lo menos, podemos decir que, así como se aplicó este método, se pudo determinar un conjunto de signos morfológicos de la lengua española, muchos de los cuales son pertinentes al nivel sintáctico. Todavía hay mucho trabajo cuantitativo pendiente, pero también y finalmente trabajo cualitativo para examinar estos signos (para agrupar los morfos en morfemas, para distinguirlos formalmente, para estudiar su morfotáctica, etc.). Seguramente, este tipo de estructura subyacente está presente en otras lenguas y su presencia en todas ellas merece ser estudiada con estas y otras herramientas que puedan concebirse para describir la unidad lingüística llamada afijo.

Capítulo 3

El clítico en el *CEMC*

Este capítulo se ocupa de los objetivos tercero y cuarto de la tesis, que son, por un lado, determinar criterios para descubrir signos gramaticales y, por el otro, aplicarlos al *CEMC* mediante la construcción de un programa computarizado. Además, se examina la hipótesis de cliticidad, según la cual esta propiedad de las palabras gráficas se puede medir mediante los índices de frecuencia, cuadros, entropía y economía. En el primer capítulo se revisaron los métodos más conocidos que se pueden aplicar en la determinación automática de vocablos gramaticales. En este capítulo se empieza con una breve reflexión sobre la naturaleza de los signos gramaticales y sobre los métodos posibles para su determinación. Después, con el objeto de examinar la hipótesis de *cliticidad*, se describe un índice cuantitativo de la cualidad que un segmento pueda tener de ser clítico como criterio para distinguir las palabras plenas de las gramaticales. Después se presenta la mecánica detrás del programa construido para seleccionar los signos más gramaticales del *CEMC* y, por último, se presentan los resultados de este procedimiento.

3.1 Sobre los signos gramaticales

En esta sección se explora la distinción entre formas gramaticales —aquellas que le dan estructura al texto o discurso— y vocablos no gramaticales, es decir, las formas léxicas o de contenido. Claro está que no podemos hablar de una frontera absoluta y fija entre estos tipos de signos, pero vale la pena investigar sus características formales. Esto es importante aquí para determinar los criterios cuantitativos que se puedan aplicar a un corpus en el descubrimiento de este tipo de signos. Para esto se examina la noción de clítico y se esboza una definición provisional, como la que se planteó para el afijo en el capítulo anterior, que nos permita relacionarlo con los signos gramaticales en general.

Como se estableció en la introducción, una de las premisas de este trabajo es la definición de palabra como una secuencia de caracteres alfabéticos comprendida entre dos espacios (véase la página 14 de la introducción). Pero como las cosas no son tan sencillas, en el capítulo sobre el afijo se tuvo que resolver el problema de los pronombres enclíticos (que en español se sufijan a la palabra gráfica) tomándolos simplemente como sufijos. Por otra parte, sobra decir que los proclíticos sí aparecen separados de las formas verbales mediante un espacio (en palabras gráficas), pero eso no los hace palabras de contenido. No todo segmento entre dos espacios puede considerarse una palabra plena y, según su carácter gramatical, difícilmente estará ligado de la misma forma a cualquier otro segmento entre espacios que aparezca a uno de sus costados.

Hay varias acepciones de los términos que se utilizan para referirse a lo que en este capítulo llamaremos clítico. Por ejemplo, en el capítulo pasado se mencionó que para Bello

los enclíticos se adhieren al final del verbo o derivado verbal, mientras que los proclíticos reciben la etiqueta de "afijos"[1]. Por otra parte, en estudios de gramática española tales como el curso de sintaxis de Gili Gaya y la gramática de Alarcos Llorach se distingue claramente entre los pronombres personales átonos o complementarios y los lugares en que ocurren con respecto al verbo: posiciones proclítica (proclisis) y enclítica (enclisis)[2]. Sin embargo, aunque el pronombre y posición no se conciben como lo mismo, hay una tendencia a considerar a los pronombres personales átonos como *los clíticos* del español.

De todas maneras, dentro de la misma tradición los clíticos también pueden ser otras cosas. Así, para Lázaro Carreter la secuencia 'a mi' del sintagma 'a mi casa' está en proclisis con respecto a 'casa'. Lo importante aquí es que la *clisis*[3] se caracteriza por la ausencia del acento prosódico en los segmentos que se adhieren al vocablo ortotónico.

Sin embargo, en el habla los patrones de acentuación cambian según las necesidades y particularidades de los hablantes y el discurso, "dependiendo del ritmo de la elocución"[4], de tal manera que no es raro que un segmento átono reciba una intensidad accesoria: 'dígamel**ó**', 'explicámel**ó**', '¿**ý** tus asuntos?'.

Además, determinar dónde están los acentos prosódicos de los vocablos en una investigación automática no es un problema menor[5]. La convención de acentuación del español sirve

[1]Con esa etiqueta Bello obviamente no se refiere a lo que investigamos como afijo en el capítulo anterior. *op. cit.* [16] 1953, pp. 120 y 286.

[2]Gili Gaya, *Curso superior de sintaxis española* [55], Vox/Bibliograf. Barcelona, 15ª ed., 1994, §177, p. 235; Alarcos Llorach, *Gramática de la lengua española* [3], Espasa Calpe. Madrid, 1999 [1994]. p. 246.

[3]Fernando Lázaro Carreter, *Diccionario de términos filológicos* [91]. Gredos. Madrid, 1990, *s.v.* CLISIS. PROCLISIS y ENCLISIS.

[4]Alarcos Llorach, *op. cit.* [3] 1999, §46 y §48, pp. 48-49.

[5]Hacerlo manualmente, además de que metodológicamente no sería compatible con lo planteado en la introducción, sería una empresa colosal, dado el tamaño del corpus.

hasta cierto punto, pero hay formas gráficas monosilábicas que, según su uso gramatical, a veces se acentúan y a veces no. Además, otras lenguas no marcan gráficamente las sílabas tónicas. Así, aunque ésta es la característica definitoria tradicional de los clíticos, aquí no podrá ser el eje de la investigación cuantitativa.

El que los segmentos sean o no átonos es una cuestión en cierta manera relacionada con su longitud. En el capítulo pasado se apuntó que los afijos tienden a ser más cortos como consecuencia del desgaste fonológico que sufren por su uso continuo, pero que no hay nada que garantice que esto ocurra instantánea ni inmediatamente después de que adquiera un carácter afijal. De hecho, aunque el uso subordinado de un segmento implique una pérdida de fuerza prosódica a corto plazo, por ahora no hay manera de medir que tan átonos son los segmentos solamente a partir del corpus. Además, si el objetivo último es *descubrir* signos gramaticales, sobra decir que éstos no son necesariamente átonos.

Por otra parte, la noción de clítico implica una complejidad que aquí no hay espacio para examinar[6]. Sin embargo, como establece Spencer, "there seems to be some consensus that it is necessary to separate out the syntactic properties of clitics from their morphophonological properties"[7], por lo que estudiar cuantitativamente las relaciones de los clíticos con los elementos que acompañan (apenas un aspecto de lo morfo-sintáctico), sin tomar en cuenta sus características fonológicas, debe ser un ejercicio interesante que, como veremos, no carece de resultados sugestivos.

---

[6]Véanse, por ejemplo, el capítulo sobre los clíticos de Spencer, *op. cit.* [129] 1991, pp. 350-391; y Halpern, "Clitics" [62] en Spencer y Zwicky, eds., *The Handbook of Morphology* [130], Basil Blackwell, Oxford, 1998, pp. 101-122.

[7]Spencer, *op. cit.* [129] 1991, p. 382.

De hecho, en sus relaciones con los elementos que los acompañan hay varias maneras en que los clíticos se parecen a los afijos: constituyen una clase relativamente pequeña y no ocurren solos, sino junto a formas de contenido (típicamente de baja frecuencia), sus significados están mucho más restringidos que los de las formas de contenido (llevan información de tipo gramatical) y tienden a ser cortos. Sin embargo, son diferentes en que son signos gramaticales muy ligados a las palabras con las que ocurren pero un poco más alejados de sus bases que los afijos (pueden incluso representarse como segmentos gráficamente independientes). En otras palabras, se adhieren al signo de contenido, pero dependen menos de éste que los afijos y, más que otros signos a su alrededor (si una raíz tiene afijos, éstos ocurren entre la raíz y los clíticos).

Con base en esto podemos hipotetizar que los otros signos gramaticales, aquellos más independientes en cuanto a los lugares en donde pueden ocurrir, también guardan cierta relación de dependencia con respecto a los signos de contenido. Es decir, la fuerza de adhesión a sus contextos, si bien menor que la de los clíticos, será mayor que la de las formas menos gramaticales y más léxicas. Tal vez sea un poco forzado referirse a la asociación de las palabras gramaticales con las de contenido como cliticidad: sencillamente no cualquier segmento de uso gramatical sería considerado por la tradición propiamente un clítico. Pero como veremos a continuación vale la pena investigar la relación entre lo que es el carácter de un clítico y el de una palabra de uso gramatical. Las formas gramaticales dan estructura y, como diría Meillet, se desgastan convirtiéndose en clíticos típicos (o desapareciendo). De alguna manera cabría suponer que las formas gramaticales, en su tendencia a asociarse con otras formas, se parecerán más a un clítico típico que a los vocablos de contenido. Entonces, consideremos

por ahora a la cliticidad y a la cualidad de ser una forma gramatical (sin ser un clítico típico del español) como dos casos de lo mismo: dos maneras de tener un carácter gramatical, una donde el segmento tiene una gran tendencia a ligarse a otros, otra donde el segmento tiene menos fuerza de adhesión, pero goza de la libertad de ocurrir en contextos más diversos.

En ese contexto, podemos esperar que además de las particularidades fónicas mencionadas arriba (atonicidad y longitud reducida por desgaste fonológico), los clíticos funcionen y se comporten morfológica y sintácticamente de manera más restringida que otros tipos de segmentos, cosa que se examinará hacia el final de este capítulo.

3.2 El clítico como pariente del afijo

En esta subsección se examina la aplicación al exterior de la palabra de los criterios formales utilizados en el capítulo anterior —específicamente la entropía, cociente de de Kock y el número de cuadros— para corroborar si pueden extenderse a una investigación de los objetos que se adhieren a las palabras. Esto, además de servir de vínculo entre éste y aquel capítulo, demuestra la pertinencia de esos criterios en el ámbito del sintagma.

No es de sorprenderse que estos mismos criterios se puedan aplicar al descubrimiento de clíticos. El esquema de la figura 3.1 representa esta idea. Allí se ilustra cómo los signos gramaticales de un nivel se adhieren a los de contenido del mismo nivel para constituir unidades del siguiente nivel, formando estructuras anidadas. Además, si hemos de concebir a los afijos y a los clíticos como dos grados diferentes de fusión de unidades originalmente independientes[8], podemos suponer que todas aquellas palabras con más función gramatical

[8]Dos estados de lo que Meillet definió como "gramaticalización", que es el proceso de evolución de las for-

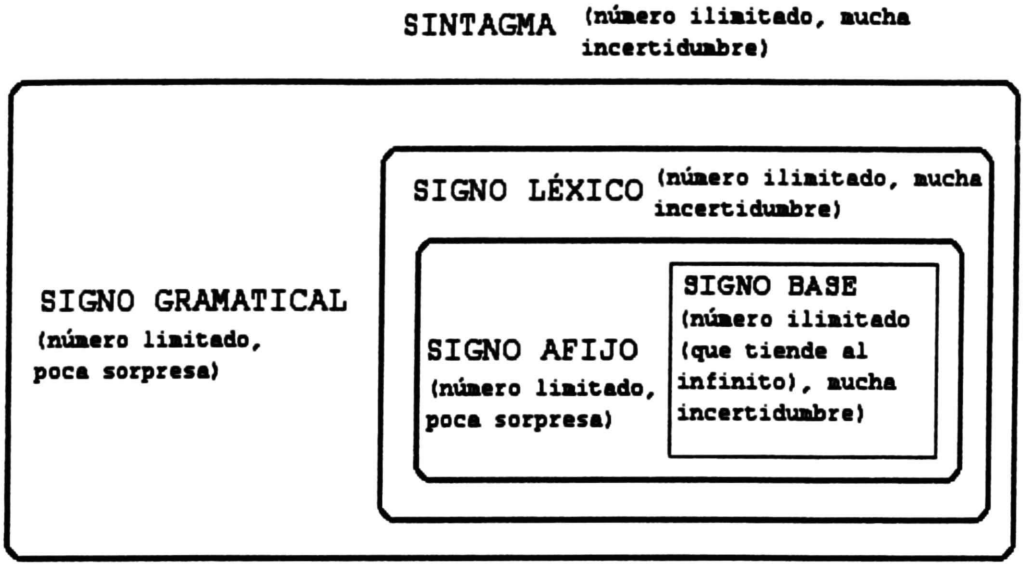

Figura 3.1: Anidamiento de signos al interior y al exterior de la palabra

que los vocablos de contenido se pueden descubrir mediante métodos similares a los del descubrimiento de afijos.

Es claro que de alguna manera las estadísticas de digramas rinden cuenta de la asociación entre vocablos de un corpus, según vimos en la sección de métodos de adquisición léxica (en lo concerniente al descubrimiento de colocaciones). Sin embargo, como vimos en el capítulo sobre el afijo, no prometen producir resultados más finos que los criterios que toman en cuenta la estructura interna del léxico y que, por lo tanto, merecen tomarse en cuenta en el cálculo de un índice de afijalidad.

Además, dado que las estadísticas de digramas miden la asociación o falta de ella entre pares de vocablos, sin especificar cuál de los dos es más dependiente de o está más aso-

mas gramaticales a partir de palabras de contenido o formas léxicas ("L'évolution des formes grammaticales" [101] en *Linguistique historique et linguistique générale* [100]. La Société de Linguistique de Paris, Paris, 1958. pp. 130-148).

ciado al otro (por lo que, en el capítulo anterior, no se tenía certeza en determinar —sin intervención del investigador— cuál de los dos segmentos era el afijo), no parecen muy atractivas en la determinación de clíticos. En comparación, los otros índices para descubrir afijos (especialmente entropía y economía) mostraron varias ventajas que vale la pena explorar: indican una dirección de "dependencia", son medidas de fenómenos lingüísticos. toman en cuenta la estructura interna del léxico completo, etc. A continuación se examinan los criterios lingüísticos que se aplicaron en el capítulo anterior para determinar su aplicabilidad al exterior de la palabra, es decir, su utilidad en el descubrimiento de clíticos.

3.2.1 Número de cuadros

Para calcular un índice de cuadros, se puede redefinir la estructura llamada cuadro como el conjunto de cuatro cadenas de caracteres delimitadas por espacios, dos de las cuales aparecieran en el corpus a la derecha de las otras dos (y viceversa. es decir, estas últimas a la izquierda de las primeras). En otras palabras, un cuadro puede constituirse mediante cuatro segmentos, s_{a_1}, s_{a_2}, s_{b_1} y s_{b_2}, cuyas combinaciones, "$s_{a_1}s_{b_1}$", "$s_{a_1}s_{b_2}$", "$s_{a_2}s_{b_1}$" y "$s_{a_2}s_{b_2}$" se deben atestiguar en el corpus sin signos de puntuación de por medio. Así. ejemplos de cuadros serían:

1. "el cuento", "el día", "este cuento", "este día";
2. "me pongo", "te pongo", "me dieron", "te dieron";
3. "cuando los", "cuando más", "aunque los", "aunque más";
4. "niños cuyos", "niños suyos", "pueblos cuyos", "pueblos suyos";
5. "de mi", "de sí", "para mi", "para sí":
6. "por la", "por su", "ante la", "ante su";

7. "se la", "se le", "te la", "te le";

8. "comisión nacional", "comisión general", "instituto nacional", "instituto general";

9. "hombre rana", "hombre araña", "niña rana", "niña araña", etc.

Sin embargo, hay varias observaciones que hacer con respecto a estas estructuras. Primero, se trata de combinaciones en las que incurren los signos del nivel léxico y podrán ser muy numerosas aunque no estén formadas por signos tan frecuentes. Nótese que cada cuadro contado para una segmentación involucra al segmento asumido (o no) como clítico y a otro semejante que alterne con él. Es decir, los cuadros mismos contienen la gama de signos del nivel siguiente producidos por este grupo de supuestos clíticos. En ese sentido, se trata también de una medida de economía de signos (como se pudo haber argumentado en el capítulo sobre el afijo). La diferencia entre esta economía y la calculada mediante el cociente de de Kock es que este último requiere que la multiplicidad de signos resultantes se correlacione con un bajo número de signos estructurales. Por otra parte, los cuadros bien pueden ser un índice de asociación entre segmentos semejantes (o los dos son o parecen clíticos, como los núms. 5, 6 y 7 de arriba; o los dos palabras plenas, como los núms. 8 y 9), es decir, de frecuencia similar y pertenecientes a conjuntos de tamaños comparables (cosa que no se reflejó en el experimento del capítulo anterior porque sólo se contaron los cuadros donde los segmentos de uno de los extremos de los vocablos guardaban una relación de mayor frecuencia y menor número que los de los otros extremos[9]) y aquí no estamos buscando determinar pares de signos semejantes, sino combinaciones de signos gramaticales con signos de contenido. Por estas razones y otras similares a la omisión de la frecuencia de los segmentos del corpus como

[9]Recuérdese también que, si bien tuvieron tanto éxito en el descubrimiento de sufijos, no lo fue así en la determinación de prefijos.

factor de la afijalidad (una fluctuación importante entre sufijos y prefijos, además de que el número de cuadros también parece ser más un resultado que una propiedad inherente de los segmentos)[10], esta medida dejó de aplicarse en este experimento.

3.2.2 Entropía

Para nuestros propósitos, el cálculo de la entropía se puede llevar a cabo construyendo una cadena de Markov[11] donde los estados no correspondan a las categorías gramaticales, es decir, donde cada transición tenga su propio estado (el caso en que cada vocablo del corpus es una categoría propia). Véase la figura 3.2, que contiene una minicadena de este tipo (la que se construyó a partir del *CEMC* consta de alrededor de 79,000 estados) que rinde cuenta de oraciones tales como "me oyó un elefante". De esta manera, para cada estado o vocablo se puede calcular una cantidad de entropía con respecto al conjunto de estados siguientes. Similarmente, se calcularon las entropías *en reversa* para cada vocablo. Véase la figura 3.3, que contiene una minicadena que rinde cuenta *en reversa* de sintagmas tales como "un sordo disparo", "este disparo sordo" y "les disparo" (nótese que se trata solamente de los significantes de los sintagmas, es decir, el hecho de que sordo pueda fungir igual como

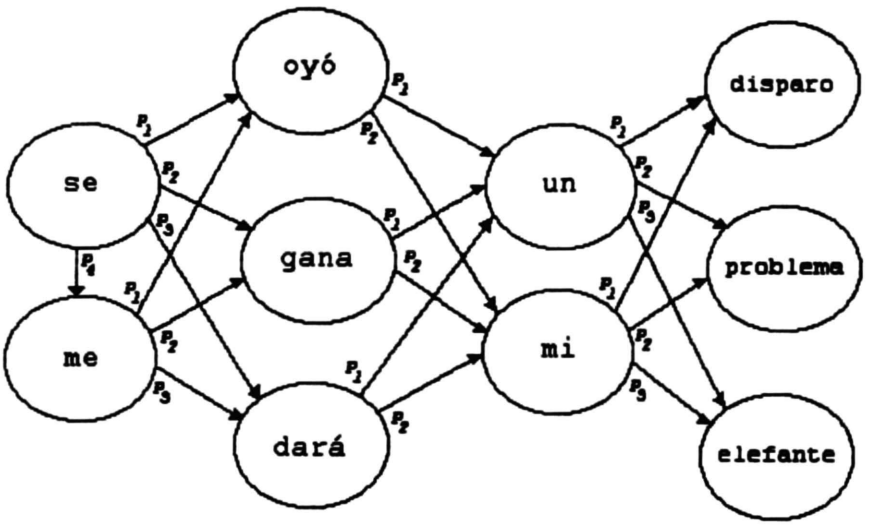

Figura 3.2: Cadena de Markov de primer orden para calcular entropías de cada vocablo

adjetivo que como sustantivo no se refleja en estas cadenas[12]).

Nótese que estos cálculos de entropía no son tan finos como los que se calcularon en el capítulo pasado. La diferencia es que allí se construyeron dos árboles, para representar las palabras a ser segmentadas, uno en cada dirección (de izquierda a derecha y de derecha a izquierda). Cada uno de estos árboles corresponde a una cadena de Markov de varios grados (al final de cada vocablo la cadena es de tantos grados como su longitud menos uno); es decir, el estado de un fonema inicial es muy diferente al estado del mismo fonema al interior

[10]También es cierto que el procedimiento para determinar estas estructuras exhaustivamente es, al igual que en el nivel morfológico, muy caro en términos de tiempo de procesamiento.

[11]Véase más adelante una definición formal de las cadenas de Markov (página 192).

[12]Esto no quiere decir que no se puedan construir cadenas que reflejen estas diferencias. De hecho, las cadenas más usadas en el análisis sintáctico probabilístico representan las palabras como arcos y sus categorías gramaticales como estados, de tal manera que una flecha representando al adjetivo 'sordo' llegaría al estado "adjetivo", mientras que otra flecha representando al sustantivo con el mismo significante apuntaría al estado "sustantivo".

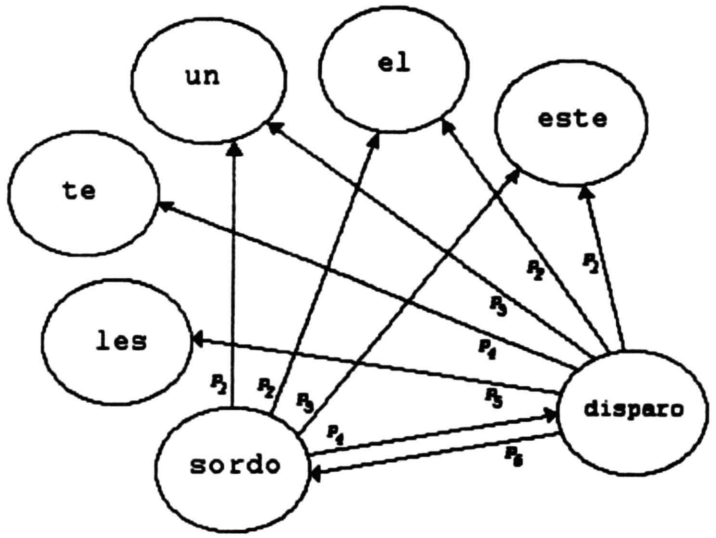

Figura 3.3: Cadena de Markov para calcular entropías *en reversa*

del vocablo. En otras palabras, en cada estado de la estructura arbórea hay una historia de estados particular al vocablo en cuestión (el estado 'o' del vocablo 'dos' no es el mismo estado 'o' del vocablo 'disparo' —véase la figura 3.4). Por otra parte, en una cadena de Markov de primer grado como la que se construyó para la investigación de este capítulo, cada forma tiene uno y sólo un estado (la forma 'sordo' corresponde a uno y sólo un estado), compárese la figura 3.4 con las figuras 3.2 y 3.3.

3.2.3 Índice de economía

El índice de economía también se calculó a partir de la cadena de Markov. Como se recuerda, nuestro índice es simplemente el cociente del total de formas que acompañan a un segmento (con las que forma sintagmas) dividido entre el número de aquellas que alternan con él (con las que forma un paradigma). El esquema markoviano nos permite tener a acceso

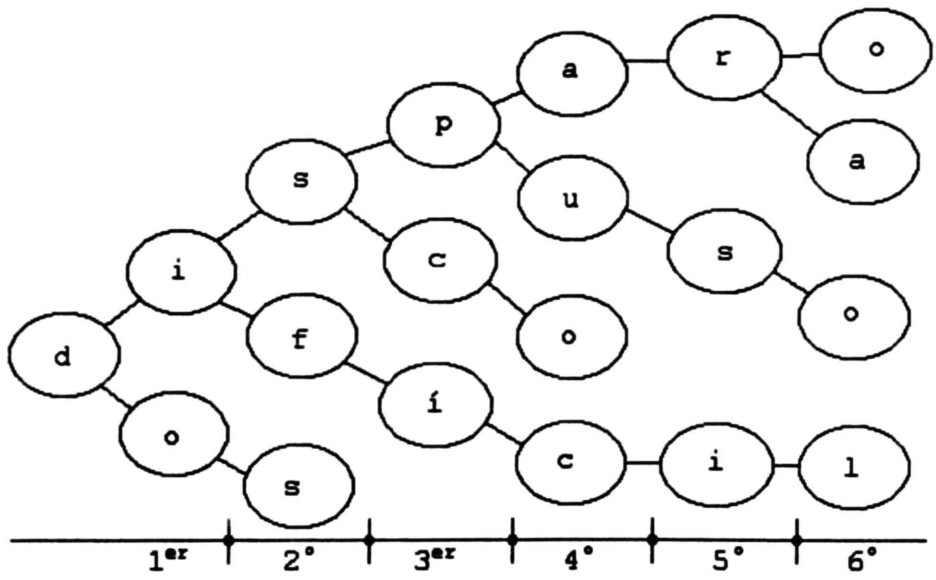

Figura 3.4: Cadena de Markov arbórea (de varios órdenes)

Este fragmento de estructura arbórea codifica algunos vocablos de izquierda a derecha. El utilizado en el capítulo anterior contiene todos los vocablos encontrados en el *CEMC* y sus frecuencias que aquí no aparecen representadas (tampoco aparecen las modificaciones de caracteres hechas para reflejar las características fonológicas del español, véase la tabla A.7 del apéndice).

a todas las palabras que ocurren a ambos lados de cada vocablo, de tal manera que la simple cuenta de todo lo que ocurre a la derecha o a la izquierda nos indica qué tan económico es el uso del segmento examinado (en comparación, por supuesto, con las cuentas de los otros segmentos). El problema es determinar los vocablos con que alterna (los que están en relación paradigmática), principalmente porque muchísimos vocablos ocurren en varios paradigmas. También hay que resaltar que los conjuntos de palabras que alternan en un "paradigma", según se puede rastrear en una cadena de Markov como la que se construyó para este experimento, no corresponden exactamente a los paradigmas de las categorías gramaticales. De hecho, cualquier cosa puede aparecer en cualquier "paradigma", a veces por error —errores de transcripción como poner "de" en lugar de "se" en "se perdió" ("de

perdió"), por lo que "de" aparecería en el paradigma de los pronombres proclíticos—, a veces por su ocurrencia en construcciones peculiares (como al transcribir un tartamudeo o alguna expresión donde se repiten ciertos segmentos "la la casa de de mi tía" que resultaría por ejemplo en "la" y "de" como miembros del paradigma nominal) y no tan peculiares (como cuando "se la comió" resulta en que "la" entre al paradigma verbal porque ocurre después de "se" que precede infinidad de formas verbales), etc. Por esta razón, se examinaron algunas alternativas a la medida de economía. El procedimiento elegido en este capítulo para calcularla se presenta más adelante (página 198).

3.3 Un índice de puntuación

Esta última subsección examina la aplicación de un índice de signos de puntuación en la determinación de fronteras sintagmáticas. Esto es pertinente, porque la variedad y abundancia de marcas de puntuación de diferentes índoles y el hecho de que su distribución no sea aleatoria, sino que muy a menudo sigue ciertos patrones asociados con la estructura sintáctica de los textos, es una invitación a examinar su utilidad para descubrir los límites entre sintagmas.

En la figura 3.5 aparece una representación esquemática de una manera de representar la información sobre la puntuación de un texto en una cadena de Markov. Normalmente cada signo de puntuación es considerado un estado en la cadena, de tal manera que, al ocurrir después de un vocablo, hay una transición del estado que representa este vocablo al estado que representa el signo de puntuación particular y luego de ese estado al estado que

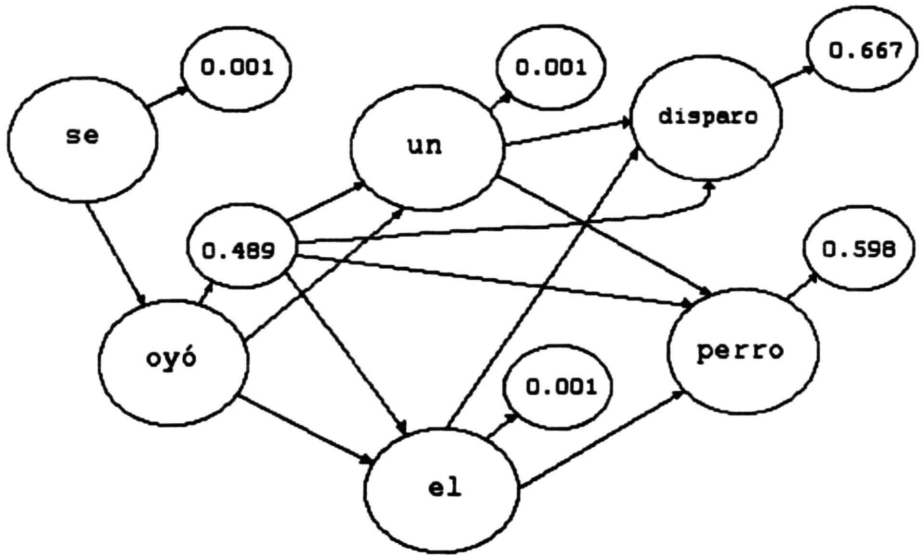

Figura 3.5: Cadena de Markov con información cuantitativa de los signos de puntuación

representa el siguiente vocablo. Pero en el esquema utilizado en este experimento (el cual aparece simplificado en la figura 3.5), hay un registro de signos de puntuación que aparecen en el corpus después de cada vocablo. Los óvalos más pequeños representan a este registro que aquí aparece como un porcentaje (la proporción de transiciones entre vocablos con un signo de puntuación de por medio). De alguna manera, esta información convierte a una cadena de Markov de primer orden en una híbrida de primer y segundo órdenes, donde para los vocablos continúa siendo del tipo original, pero para éstos y su relación con la puntuación es ahora del segundo orden. Así, un signo de puntuación dado representa un estado particular de la ocurrencia de un vocablo (porque puede ocurrir ya sea directamente o con algún tipo de marca entre éste y el próximo vocablo). Nótese cómo el porcentaje hipotético de signos de puntuación que siguen a pronombres como "se" y a artículos como "un" y "el" debe ser muy bajo (de hecho, cero en córpora pequeños o completamente "*gramaticales*") en comparación

con la proporción de marcas que posiblemente ocurren después de formas verbales y nominales como "oyó", "disparo" y "perro", de las cuales no sería extraño esperar un 50%.

Esta simple diferencia distribucional, que en el *CEMC* es más que una intuición, es un indicio de fronteras de sintagmas digno de tomarse en cuenta. De hecho, un alto grado de cliticidad asociado a un vocablo debe oponerse a un porcentaje alto de ocurrencias junto a ese vocablo de signos de puntuación (precisamente del lado donde la cliticidad es más alta). Es decir, si la cliticidad es alta, hay poca puntuación. Si hay mucha puntuación, la cliticidad debe ser baja.

Sin embargo, también vale la pena considerar las diferencias entre los signos de puntuación, porque no todos *separan* igual. Es decir, algunos deben *pesar* más que los otros. Ignorando los múltiples errores de puntuación que inevitablemente se encuentran en un texto como el *CEMC*, en español, como en otras lenguas de origen europeo, el punto sirve para separar períodos, los que pueden estar formados tanto de frases como de oraciones. De manera similar, los dos puntos y el punto y coma marcan las fronteras entre estructuras similares. La coma, sin embargo, puede ocurrir al interior de diversos tipos de sintagmas (por ej. separando adjetivos en una frase nominal o adverbios en una construcción adverbial). Los signos de interrogación, de admiración y parentéticos (guiones, llaves, paréntesis, etc.) se utilizan prototípicamente para separar períodos, ya sean oraciones o frases, pero en los hechos pueden aparecer, aunque con baja frecuencia, en cualquier lado. Además, tampoco el punto es un delimitador de sintagmas tan confiable como podría esperarse, porque tiene otras funciones, como la de cerrar abreviaturas, cosa que no necesariamente coincide con las fronteras sintagmáticas (por ej. *Le dijo a la Sra. Ramírez que viniera*).

De allí la idea de asignarle a cada signo de puntuación un peso de acuerdo a su confiabilidad como delimitador. A reserva de hacer una verdadera investigación de la puntuación —cosa que queda fuera del ámbito de esta tesis—, se fijaron intuitivamente ciertos valores a la ocurrencia de cada signo de puntuación. En la tabla 3.1 se especifican estos valores. Nótese

Tabla 3.1: Pesos asignados a los signos de puntuación

signo	valor
: :	5
.	3
,	1
¿ ? ¡ !	1
( ) { } —	1
...	1

que al punto no se le asignó un valor tan alto como al punto y coma y a los dos puntos. La razón de eso es, por un lado, que el punto es más ambiguo (incluso que la coma) al servir también como marca de abreviaturas (véase la estrategia para determinar abreviaturas en el apéndice dedicado al *CEMC*, página 360). Por otro lado, tanto el punto y coma como los dos puntos están entre los caracteres menos ambiguos del corpus (véanse las tablas A.2 y A.3 en el mismo apéndice). El caso es que, al recorrer el corpus para construir la cadena de Markov, cada vez que ocurrió uno de estos signos, el valor correspondiente se sumó al registro de la transición entre el vocablo inmediatamente anterior y el inmediatamente posterior.

3.4 Hacia un índice de *cliticidad*

En este apartado se describe un índice cuantitativo de *cliticidad* como criterio para distinguir los clíticos de las palabras plenas, es decir, se propone una medida del carácter de clítico que pueda tener un segmento dado (medida de alguna manera también aplicable en

el reconocimiento de palabras gramaticales no necesariamente clíticas). Como quedó establecido en el capítulo anterior, la notable ventaja de las medidas de economía y entropía (y del número de cuadros), en la predicción de fronteras morfológicas, se explica en el hecho de que éstas corresponden a nuestra concepción tradicional de cómo son los afijos. Sin embargo, como se apuntó arriba, estas medidas también se pueden calcular hacia el exterior de la palabra, con el objeto de determinar las formas más clíticas de entre el conjunto de vocablos (palabras gráficas) extraído de un corpus, ya que, como vimos también arriba, se trata de relativamente pocas formas muy frecuentes que contienen menos información que los vocablos más plenos.

En la introducción me referí a esta cualidad como la *cliticidad* de un segmento s_x y —de manera similar a la afijalidad— se propuso una hipótesis para medirla:

$$CL(s_x) = \frac{f_x c_x k_x}{h_x}$$

donde f_x, c_x, k_x y h_x representan respectivamente la frecuencia, la cuenta de cuadros, la economía y la entropía del segmento s_x, según se puedan estimar a partir de un corpus Ψ. La similitud con el fenómeno de afijalidad nos permite considerar la eliminación de la frecuencia como parámetro pertinente y la elección de la entropía mayor (en vez de la menor) como marca de frontera de signos. De esta manera, nos quedamos con la fórmula:

$$CL(s_x) = k_x c_x h_x$$

Sin embargo, también aquí hay que reconsiderar un par de cosas. Primero, el número de cuadros parece ser (como ya se mencionó arriba), al igual que la frecuencia, una manifestación o resultado de las otras dos magnitudes, especialmente de la economía. De hecho, el número

de cuadros puede considerarse en sí mismo otra medida de economía, porque una mayor presencia de este tipo de estructuras es sólo posible con un número mayor de signos producidos en el nivel siguiente[13]. Entonces, se puede hipotetizar que la cliticidad se puede medir de la siguiente manera:

$$CL(s_x) = k_x h_x \tag{3.1}$$

Esto es, la cualidad de s_x de ser un clítico es directamente proporcional al producto de alguna medida de economía (k) por alguna medida del contenido de información (h) que cabe esperar cuando aparece en la cadena hablada. Todas estas cantidades se pueden calcular, por ejemplo, mediante una cadena de Markov (de uno o, si acaso lo permite la capacidad del equipo utilizado, varios grados) construida a partir de un corpus Ψ.

La segunda consideración de importancia es el fenómeno de la puntuación que, como se mostró también arriba —si bien particular a la lengua escrita, pero también inevitable para representar la lengua hablada mediante caracteres—, constituye un criterio adicional muy útil en la determinación de asociación de clíticos y palabras plenas y, por ende, en el descubrimiento de fronteras sintagmáticas.

De allí la posibilidad de concebir dos tipos de cliticidad, una general para transcripciones de lengua hablada, como la descrita en la ecuación 3.1, y otra para la lengua escrita que tome en cuenta la puntuación:

$$CL_{esc}(s_x) = \frac{CL(s_x)}{r_x} = \frac{k_x h_x}{r_x}$$

donde r_x es el número de signos de puntuación que ocurren entre el segmento s_x y los vocablos plenos a los que supuestamente se adhiere. La presencia de numerosas marcas no

[13]De hecho, el número de cuadros (aunque requiere de córpora muy grandes) parece más adecuado para medir la economía de signos en fenómenos de composición, no de afijación ni de cliticización.

alfabéticas entre los caracteres que constituyen las palabras gráficas implica cierta resistencia (r) a la fuerza de asociación entre clíticos y vocablos plenos, por lo que debe ser inversamente proporcional a la cliticidad.

Pero aparte de las virtudes de combinar estas medidas, hay que tomar en cuenta la cuestión de la normalización (esto es, en el intervalo [0,1]). Hay varias maneras de proceder para normalizar nuestro índice de cliticidad, pero aquí se eligieron fórmulas semejantes a la utilizada en el capítulo anterior (para evadir los problemas que surgen al considerar la combinación de estos índices en cuanto a la naturaleza de sus magnitudes, escalas y unidades):

$$CL^n(s_x) = \frac{\frac{h_x}{\max h_i} + \frac{k_x}{\max k_i}}{2} \tag{3.2}$$

$$CL^n_{esc}(s_x) = \frac{CL^n(s_x) + (1 - \frac{r_x}{f_x})}{2} \tag{3.3}$$

$$CL^n_{esc}(s_x) = \frac{\frac{h_x}{\max h_i} + \frac{k_x}{\max k_i} + (1 - \frac{r_x}{f_x})}{3} \tag{3.4}$$

donde $0 \leq \frac{r_x}{f_x} \leq 1$ es la probabilidad de que ocurra un signo de puntuación entre el segmento s_x y los vocablos a los que supuestamente aparece adherido:

$$\frac{r_x}{f_x} = \frac{\text{total de marcas de puntuación que aparecen entre } s_x \text{ y los vocablos a los que supuestamente se adhiere}}{\text{total de ocurrencias del segmento } s_x}$$

De la misma manera en que —para calcular una medida de afijalidad normalizada— promediamos en el capítulo anterior los índices normalizados para evitar números muy pequeños al multiplicar, aquí se opta por substraer esta probabilidad de la unidad (en lugar de dividir) para evitar resultados mayores a uno. En otras palabras, la medida de puntuación calculada para este experimento es simplemente el complemento de la probabilidad de ocurrencia de signos de puntuación[14] inmediatamente después (o antes) del segmento examinado.

[14] El número de marcas de puntuación que ocurrieron entre dos segmentos dividido entre el número total de veces que ocurrieron los segmentos en esa secuencia (separados o no por signos de puntuación).

3.5 Cuestiones preliminares en la determinación de los clíticos del *CEMC*

En esta sección se describe la mecánica detrás del programa construido para seleccionar las palabras más gramaticales del *CEMC*. En la primera parte se presentan las nociones formales preliminares relativas a las cadenas de Markov. En la segunda se describe la manera en que se calcularon la entropía y la economía a partir de éstas.

3.5.1 Las cadenas de Markov

Esta subsección presenta una formalización posible del esquema markoviano. Primero que nada, es necesario enfatizar que la pertinencia de este esquema para la lingüística no reside en que sea el esquema descriptivo más natural a las lenguas naturales, sino que se trata de una herramienta muy útil en el cálculo de los índices de entropía y economía, abstracciones que tampoco son los mejores ni los únicos esquemas descriptivos del lenguaje, sino los mejores medios encontrados aquí para estimar ciertas propiedades lingüísticas de los objetos de un corpus.

Las cadenas de Markov se utilizan para representar secuencias de variables aleatorias (o símbolos) cuyas ocurrencias no son independientes[15], es decir, cuando el valor de cada una depende de los objetos previos en la secuencia.

En el capítulo anterior definimos a nuestro corpus Ψ como una secuencia de ocurrencias de palabras de tamaño ξ: $\Psi = o_1, o_2, o_3 \ldots, o_\xi$, y a V como el conjunto de vocablos

[15]Manning y Schütze, *op. cit.* [93] 1999, pp. 318-320.

$\{v_1, v_2, v_3, \ldots, v_\Omega\}$. Entonces, podemos considerar a Ψ como la secuencia de variables aleatorias cuyos valores posibles están en el conjunto V (el conjunto de estados en la cadena). Así, la probabilidad de que una palabra o_t sea una instancia del vocablo v_i (es decir, que o_t tome el valor v_i, lo que es lo mismo que llegar al estado v_i) depende de los valores previos que tomaron las ocurrencias inmediatamente anteriores (de los estados a los que se arribó previamente). De hecho, al haber un horizonte limitado de estados previos, se puede asumir —dependiendo del orden del esquema (primero, segundo, etc.)— que la probabilidad del estado presente, dado un número limitado de estados previos, es igual a la probabilidad del mismo, dada la secuencia completa de estados previos en Ψ. Por ejemplo, para una cadena de primer orden, consideraremos que:

$$P(o_{t+1} = v_j | o_1, o_2, o_3, \ldots, o_t) \approxeq P(o_{t+1} = v_j | o_t)$$

Las probabilidades de las transiciones de estado a estado en una cadena markoviana se suelen representar mediante matrices. Por ejemplo, sea $A = a_{ij} = P(o_{t+1} = v_j | o_t = v_i)$ la matriz de probabilidades:

$$a_{ij} = \begin{cases} P(o_{t+1} = v_1 | o_t = v_1) & P(o_{t+1} = v_2 | o_t = v_1) & P(o_{t+1} = v_3 | o_t = v_1) & \cdots & P(o_{t+1} = v_\Omega | o_t = v_1) \\ P(o_{t+1} = v_1 | o_t = v_2) & P(o_{t+1} = v_2 | o_t = v_2) & P(o_{t+1} = v_3 | o_t = v_2) & \cdots & P(o_{t+1} = v_\Omega | o_t = v_2) \\ P(o_{t+1} = v_1 | o_t = v_3) & P(o_{t+1} = v_2 | o_t = v_3) & P(o_{t+1} = v_3 | o_t = v_3) & \cdots & P(o_{t+1} = v_\Omega | o_t = v_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P(o_{t+1} = v_1 | o_t = v_\Omega) & P(o_{t+1} = v_2 | o_t = v_\Omega) & P(o_{t+1} = v_3 | o_t = v_\Omega) & \cdots & P(o_{t+1} = v_\Omega | o_t = v_\Omega) \end{cases}$$

donde i y j son índices que van de 1 al número total de estados que es el número de vocablos Ω. De esta manera, cada a_{ij} de la matriz A contiene la probabilidad de llegar al estado o vocablo j, una vez que se ha llegado al estado o vocablo i. Eso quiere decir que la suma de las transiciones por renglón (las probabilidades de llegar a un estado cualquiera a partir de

uno en particular) suman uno[16]: $\sum_{j=1}^{\Omega} a_{ij} = 1$.

3.5.2 Procedimiento y discusión

En este subapartado se describe y comenta el procedimiento aplicado al corpus para calcular la cliticidad de sus vocablos. Concretamente, se trata de examinar la manera en que se estimaron los índices pertinentes. Primero se describen las rutinas y las estructuras de información que se hicieron operativas. Luego se examina la manera en que se estimaron los índices mediante éstas.

Las rutinas que se construyeron para este experimento son parecidas a las del capítulo anterior. De hecho, el preprocesamiento del corpus es el mismo y los procedimientos de construcción de estructuras arbóreas para almacenar los vocablos y sus frecuencias también se aprovecharon aquí.

Por otra parte, el fichador fue modificado, primero, para registrar la ocurrencia de los signos de puntuación (tomando en cuenta los pesos intuitivamente asociados a cada signo, como se especifica en la tabla 3.1 de la página 188), segundo, para filtrar la ocurrencia de abreviaturas (con base en la lista cuya determinación está descrita, como ya se dijo, en el apéndice) y, por último, para detectar el tipo de contextos donde aparecen los números (ya sea que se trate de fórmulas, fechas o cifras monetarias).

---

[16]Para los experimentos llevados a cabo en este capítulo, una matriz de este tipo sería extremadamente "escasa", es decir, muchas celdas de la matriz tendrían una probabilidad de cero, ya que la mayoría de los vocablos son poco frecuentes (aproximadamente la mitad ocurre sólo una vez, como es común en todos los córpora) y tienden a aparecer rodeados de unos pocos (los segmentos gramaticales y las palabras de contenido más frecuentes). De esta manera, cada transición de la cadena markoviana que se construyó para este experimento se representó mediante, no la celda de una matriz, sino mediante una estructura dinámica creada sólo después de verificar la existencia de una transición en el corpus.

Además, se construyeron, como ya se dijo, dos cadenas de Markov, ambas de un grado, para almacenar las ocurrencias y frecuencias de cada par de palabras gráficas atestiguado en el corpus: una cadena con transiciones de cada palabra a todas aquellas que le sucedieron y otra con transiciones entre cada una y todas las que le precedieron.

Las cadenas se representan por lo menos con dos tipos de nodos. El primero para los estados, que aquí corresponden a los vocablos y que contiene campos que registran diversos valores (por ej. frecuencia, valor acumulado de ocurrencias de signos de puntuación, etc.). El segundo tipo de nodos es para las transiciones y tiene campos que registran los estados entre los que existe una transición y la frecuencia de dicha transición en el corpus. En otras palabras, los nodos de transición representan la ocurrencia de dos estados, es decir, de dos vocablos que aparecen adyacentes en el corpus. Si dos vocablos nunca aparecieron yuxtapuestos, no hay nodo de transición entre sus nodos correspondientes. Si ocurrieron juntos cuatrocientas o sólo tres veces en el corpus, entonces el nodo de transición lo registra en su campo de frecuencia.

En los nodos para los estados se incluyó un campo que contiene la localización de otro estado (esto es, que *apunta* a otro nodo de otro estado), cosa que permite encadenar estos nodos según distintos criterios de ordenamiento. Concretamente, al comparar automáticamente todos (o algunos de) los valores de todos los nodos se pueden ordenar de mayores a menores o vice versa (también tomando en cuenta otras características de los vocablos, como su longitud o su lugar en una secuencia alfabética, etc.). La ventaja de este esquema es que para ordenar los nodos no es necesario cambiarlos de lugar, sino sólo modificar los campos que apuntan a otros estados. Luego la cadena se recorre nodo por nodo (siguiendo el encadenamiento) para

obtener listas de vocablos ordenados según los criterios escogidos (los que se examinaron en este experimento se explicarán en la sección 3.6, a partir de la página 199).

Los procedimientos para estimar la entropía y economía de cada transición son similares a los del capítulo pasado. De todas maneras, se examinan a continuación las particularidades de su cálculo en este experimento.

Como ya se ha dicho varias veces, la entropía se estima a partir de un estado dado con respecto al conjunto de posibles estados siguientes (o precedentes). Así, el ejemplo de la tabla 2.2 (página 108) representa la entropía (o incertidumbre de lo que sigue) después de que ocurre el grafema 'p' al principio de una palabra en un texto del español de México (o el fonema /p/ en la cadena hablada). Este cálculo de entropía es natural al esquema markoviano que, como ya vimos, no es otra cosa que una red de estados posibles de un sistema, cuyas probabilidades de transición están especificadas.

De manera similar a su aplicación al interior de la palabra, la entropía entre la ocurrencia de un clítico (de hecho, de cualquier tipo de segmento) se estima calculando las probabilidades de los vocablos cuya ocurrencia inmediatamente posterior a ese segmento está documentada en el corpus y aplicando la fórmula conocida ($H(X) = -\sum_{i=1}^{n} p_i \times log_2(p_i)$).

En la tabla 3.2 aparecen las cuentas del cálculo de entropía después del estado representado por el segmento 'de' de la cadena de Markov construida para este capítulo. La medida de entropía que cabe esperar después de este segmento es simplemente la suma de los valores de la última columna. Las entropías *en reversa* se estimaron de manera similar, pero tomando en cuenta las formas anteriores.

Tabla 3.2: Entropía que cabe esperar después del segmento 'de'

	$a_{i,j}$ $B_{i,j}$	formas	p($b_{k,j}$ | $a_{i,j}$= de)	-p×log(p)
1.	DE LA	15368	0.134399	0.269731
2.	DE LOS	6845	0.0598622	0.168555
3.	DE LAS	4334	0.0379025	0.124045
4.	DE QUE	2343	0.0204904	0.0796627
5.	DE UN	2265	0.0198083	0.0776813
6.	DE SU	1978	0.0172984	0.070182
7.	DE UNA	1880	0.0164413	0.0675403
8.	DE SUS	1085	0.00948874	0.0441952
9.	DE LO	708	0.00619173	0.0314821
10.	DE ESTE	645	0.00564077	0.0292064
11.	DE MI	620	0.00542214	0.0282887
12.	DE ESTA	597	0.005221	0.0274367
13.	DE AQUÍ	535	0.00467878	0.0251003
14.	DE SER	458	0.00400539	0.0221102
15.	DE ACUERDO	439	0.00383922	0.0213556
16.	DE TODOS	391	0.00341945	0.0194166
17.	DE AGUA	387	0.00338446	0.0192527
18.	DE TODO	385	0.00336697	0.0191707
19.	DE ELLOS	324	0.00283351	0.016622
20.	DE AHÍ	309	0.00270232	0.0159806
21.	DE CADA	286	0.00250118	0.0149846
22.	DE ESA	281	0.00245745	0.0147659
23.	DE ÉL	279	0.00243996	0.0146783
24.	DE ELLA	267	0.00233502	0.0141496
25.	DE ESE	266	0.00232627	0.0141053
26.	DE TRABAJO	255	0.00223007	0.0136162
27.	DE DOS	254	0.00222133	0.0135715
	⋮	⋮	⋮	⋮
5356.	DE BISTECES	1	0.00000874539	0.000101857
	total	107887	1.0	5.89045 bits

Si bien por un lado el método para estimar la entropía no varió mucho con respecto al aplicado en el capítulo anterior, por el otro, para la estimación del índice de economía, sí hubo cambios importantes debido, primero, a la poca seguridad de que los estados que sigan o precedan al estado examinado pertenezcan (como se apuntó arriba) a uno o a unos pocos paradigmas y, segundo, a la necesidad de encontrar maneras de agilizar el procedimiento. De esta manera, se éxploraron varias alternativas que no examinaran todos los posibles segmentos en relación sintagmática con todos los posibles alternantes del segmento examinado, sino sólo la muestra de los acompañantes de éste último (los que están con él en relación sintagmática).

Es decir, se asumió por comodidad que cada forma alternaba sólo con la ausencia de sí misma. Esto no es, por supuesto, cierto. Pero fue un pretexto para explorar otras posibilidades de estimación de la economía de un segmento que, como se verá, dio buenos resultados.

Así, después de observar el comportamiento de varias alternativas (desde tomar simplemente la cuenta de segmentos acompañantes, hasta tomar en cuenta las frecuencias tanto del segmento examinado como de los acompañantes), se optó por el siguiente índice de economía para el lado derecho del segmento s_x:

$$k^{der}(s_x) = \frac{|B_x|}{\frac{\sum_{i=1}^{|B_x|} fr(b_i)}{|B_x|}} = \frac{|B_x|^2}{\sum_{i=1}^{|B_x|} fr(b_i)}, \quad \text{donde cada} \quad b_i \in B_x \tag{3.5}$$

donde $|B_x|$ es el número de vocablos que ocurren en el corpus después del segmento s_x dividido entre el promedio de ocurrencias en los vocablos en ese conjunto. Así, para que $k^{der}(s_x)$ sea grande, el segmento s_x tiene que estar seguido de muchos vocablos (b_i), cuyas frecuencias deben ser relativamente pequeñas (a mayor promedio de frecuencias menor economía). Similarmente, para el lado izquierdo del mismo segmento:

$$k^{izq}(s_x) = \frac{|A_x|}{\frac{\sum_{i=1}^{|A_x|} fr(a_i)}{|A_x|}} = \frac{|A_x|^2}{\sum_{i=1}^{|A_x|} fr(a_i)}, \quad \text{donde cada} \quad a_i \in A_x \tag{3.6}$$

donde $|A_x|$ es el número de vocablos que ocurren en el corpus antes del segmento s_x dividido entre el promedio de ocurrencias por vocablo en ese conjunto. Así, para que $k^{izq}(s_x)$ sea grande, el segmento s_x tiene que estar precedido de muchos vocablos o estados (a_i), cuyas frecuencias deben ser relativamente pequeñas (a mayor promedio de frecuencias menor economía).

Por ejemplo, aquí se puede estimar el índice de economía para el lado derecho del segmento 'de'. En la tabla 3.2 se establece que el número de vocablos que aparecen después

de este segmento es $|B_x| = 5356$ mientras que la suma de las frecuencias de todos éstos es $\sum_{i=1}^{|B_x|} fr(b_i) = 107887$. Así, la medida de la economía sin normalizar inherente al lado derecho de la preposición 'de' será, en esta investigación:

$$k^{der}(\text{de}) = \frac{5356^2}{107887} = 265.89613$$

De esta manera, el procedimiento consistió sencillamente, primero, en construir una cadena de Markov (donde cada vocablo es un estado y cada transición de estado a estado lleva asociada una probabilidad) y, segundo, en calcular para todos y cada uno de los estados índices normalizados de economía, entropía y puntuación con las fórmulas descritas arriba (aplicadas a ambos lados). Por último, con estos datos se estimaron para cada vocablo índices normalizados de cliticidad tanto general como escrita.

3.6 Resultados del procedimiento

En este apartado se presentan los resultados de la aplicación al corpus del procedimiento de descubrimiento de clíticos descrito en los apartados anteriores. Primero se examinan las formas con más cliticidad (preformas y postformas). Luego aquellas cuya cliticidad se concentra en una dirección (los clíticos propios) o en ninguna de las dos (nexos). También se presentan aquellas con más cliticidad total (las formas más "gramaticales").

Las tablas 3.3 y 3.4 muestran los resultados de la aplicación de este método al descubrimiento de formas gramaticales[17]. La primera muestra los segmentos que tienden a adherirse a los palabras que preceden (preformas), mientras que la segunda la lista de segmentos que

[17]Para todos los datos presentados en esta sección, no se tomaron en cuenta segmentos con un índice de cliticidad menor a 0.2.

tienden a ocurrir junto con las palabras que siguen (postformas). En las dos tablas se puede ver que al tomar en cuenta la ausencia de puntuación entre los segmentos, se incrementa muchísimo su carácter de clítico. Así, las últimas columnas muestran dos índices normalizados de cliticidad. La última columna es sencillamente el promedio de la cliticidad general y el índice de no puntuación ($1 - \frac{r_x}{f_x}$) asociado al segmento (tercera columna)[18].

Nótese que no se trata exclusivamente de los clíticos tradicionales (los pronombres clíticos se encuentran entre las primeras 27 preformas), sino de palabras de carácter muy gramatical. De hecho, muchas de ellas aparecen en las dos tablas, es decir. son formas que se adhieren tanto a las palabras que siguen como a las que preceden. Lo importante es observar la variedad de segmentos en las dos tablas. Además de los pronombres clíticos, aparecen preposiciones, artículos, adjetivos demostrativos. adverbios, pronombres y conjunciones. Es de notarse que los artículos y demostrativos están, como los pronombres clíticos, mucho más concentrados entre las preformas, mientras que las preposiciones ocurren, aunque repartidas en la tabla 3.3 (7 segmentos), concentradas al principio de las postformas (19 formas). Otros segmentos interesantes representan los verbos 'ser', 'estar', 'haber', 'poder'. 'deber', 'ir', 'tener' y 'hacer' (tabla 3.3: núms. 23, 29, 31, 33, 35. 38, 43, 47, 50, 53, 56, 59-61, 67, 68, 70, 71, 77, 81, 85-88, 91, 92 y 100; tabla 3.4: 20, 28, 46, 59, 60, 67, 68, 75, 78, 82, 87 y 95).

Los clíticos verdaderos se asocian más a un lado que al otro, por lo que se obtienen comparando la cliticidad de un lado con la del otro. Mientras más grande la diferencia, mayor su asociación unilateral y por lo tanto mayor su condición de clítico verdadero. Las

[18]En el resto de las tablas y de la discusión de este apartado no se vuelve a abundar en la cliticidad escrita que, como se ve, es un buen auxiliar para afinar el índice de cliticidad, pero presupone los procesos de transcripción. Me parece que antes de hacer intervenir los fenómenos de lengua escrita, conviene observar el funcionamiento de la cliticidad como efecto de los índices de entropía y economía que se aplicaron en el capítulo pasado.

Tabla 3.3: Preformas gramaticales del *CEMC* en orden de *cliticidad*

	clítico	fr.	punt.	entrop.	econ.	cliticidad "general"	cliticidad "escrita"
1.	se	33623	0.9977	1	0.9724	0.9862	0.99
2.	la	73110	0.995	0.9907	0.858	0.9243	0.9479
3.	de	114346	0.9948	0.8295	1	0.9148	0.9414
4.	un	19765	0.9923	0.9922	0.7568	0.8745	0.9138
5.	los	31231	0.9947	0.9871	0.7511	0.8691	0.9109
6.	el	51708	0.9934	0.9732	0.7459	0.8595	0.9042
7.	su	12520	0.9949	0.991	0.7166	0.8538	0.9008
8.	y	60303	0.9782	0.8829	0.8128	0.8478	0.8913
9.	una	16473	0.9878	0.9692	0.7033	0.8362	0.8868
10.	las	20882	0.9967	0.9554	0.5087	0.732	0.8202
11.	al	11179	0.9948	0.9275	0.4658	0.6966	0.796
12.	sus	5267	0.9947	0.9386	0.451	0.6948	0.7948
13.	me	10410	0.9964	0.9179	0.4597	0.6888	0.7913
14.	que	67243	0.9757	0.8336	0.4998	0.6667	0.7697
15.	del	18621	0.9934	0.8493	0.4664	0.6578	0.7697
16.	le	8502	0.9966	0.8812	0.3528	0.617	0.7435
17.	más	9778	0.934	0.8498	0.3757	0.6127	0.7198
18.	nos	3399	0.9944	0.9031	0.2923	0.5977	0.7299
19.	muy	4915	0.9927	0.8541	0.2975	0.5758	0.7148
20.	para	14655	0.9943	0.8131	0.3115	0.5623	0.7063
21.	lo	13705	0.9977	0.7626	0.3284	0.5455	0.6963
22.	te	3389	0.9991	0.8518	0.2308	0.5413	0.6939
23.	ser	2830	0.947	0.8601	0.1922	0.5262	0.6664
24.	esta	2925	0.9918	0.8429	0.2093	0.5261	0.6813
25.	tan	1591	0.9887	0.8663	0.1826	0.5245	0.6792
26.	dos	3216	0.9229	0.8124	0.2248	0.5186	0.6534
27.	les	2021	0.9985	0.8408	0.1905	0.5156	0.6766
28.	esa	1666	0.982	0.8218	0.1653	0.4936	0.6564
29.	está	4108	0.9679	0.7894	0.1933	0.4914	0.6502
30.	o	8264	0.9737	0.8319	0.1505	0.4912	0.652
31.	son	4447	0.94	0.7856	0.1889	0.4872	0.6382
32.	como	11088	0.984	0.7884	0.1694	0.4789	0.6473
33.	están	1561	0.9795	0.7983	0.1558	0.477	0.6445

tablas 3.5 y 3.6 contienen las formas del corpus cuya diferencia entre cliticidades fue mayor[19]. Están dispuestas según un índice de ordenamiento (véanse las últimas columnas), calculado mediante el producto de la menor de las cliticidades y el valor absoluto de la diferencia (para eliminar segmentos con valores de cliticidad bajos): $min(CL(s_i)) \times (|max(CL(s_i)) -$

Tabla 3.3 (continuación):
Preformas gramaticales del *CEMC* en orden de *cliticidad*

	clítico	fr.	punt.	entrop.	econ.	cliticidad "general"	cliticidad "escrita"
34.	a	45525	0.9778	0.7325	0.22	0.4763	0.6434
35.	había	2023	0.9842	0.7948	0.1517	0.4733	0.6436
36.	mi	4840	0.9957	0.7696	0.1694	0.4695	0.6449
37.	ese	1977	0.9757	0.7938	0.1447	0.4693	0.6381
38.	puede	2480	0.9782	0.7664	0.1678	0.4671	0.6375
39.	con	18747	0.9963	0.7149	0.2069	0.4609	0.6393
40.	no	31676	0.7875	0.6985	0.2217	0.4601	0.5693
41.	sin	3179	0.9994	0.7488	0.1701	0.4595	0.6394
42.	tu	1278	0.9961	0.7775	0.1184	0.448	0.6307
43.	ha	4105	0.9803	0.7337	0.1388	0.4363	0.6176
44.	es	18601	0.9698	0.708	0.1639	0.4359	0.6139
45.	estas	765	0.9922	0.7704	0.1006	0.4355	0.6211
46.	otros	1499	0.8933	0.7506	0.1174	0.434	0.5871
47.	han	1947	0.9974	0.7383	0.1283	0.4333	0.6213
48.	esos	741	0.9757	0.7629	0.08994	0.4264	0.6095
49.	estos	1005	0.9861	0.7575	0.09308	0.4253	0.6122
50.	pueden	944	0.9788	0.7416	0.101	0.4213	0.6071
51.	e	1325	0.7464	0.6864	0.138	0.4122	0.5236
52.	nuestra	768	0.9674	0.7327	0.09033	0.4115	0.5968
53.	fue	2827	0.9653	0.7338	0.0857	0.4097	0.5949
54.	tres	1598	0.9224	0.7181	0.09979	0.409	0.5801
55.	grandes	845	0.8757	0.7199	0.09797	0.4089	0.5645
56.	hay	4013	0.9594	0.7188	0.09869	0.4087	0.5923
57.	mis	963	0.9969	0.7193	0.09251	0.4059	0.6029
58.	cuya	267	0.9963	0.7333	0.07668	0.405	0.6021
59.	estaba	1368	0.973	0.7331	0.07591	0.4045	0.594
60.	debe	1318	0.9863	0.712	0.09445	0.4032	0.5976
61.	fueron	861	0.9384	0.7426	0.06333	0.403	0.5815
62.	por	19835	0.996	0.6882	0.1141	0.4012	0.5994
63.	otras	1071	0.9262	0.7108	0.08851	0.3997	0.5752
64.	cuando	4765	0.9815	0.7001	0.09889	0.3995	0.5935
65.	unas	643	0.9813	0.7274	0.0637	0.3955	0.5908
66.	ni	2392	0.9946	0.7369	0.05153	0.3942	0.5943

$min(CL(s_i))|$). La primera tabla (3.5) presenta los proclíticos y la segunda (3.6) los enclíticos.

Por ejemplo, compárense los valores del segmento 'y' que aparecen en las tablas 3.3 y 3.4. Al restarlos se obtiene un valor cercano a 0.1 ($|0.8478 - 0.9515| = |-0.1037| = 0.1037$). También obsérvense los del clítico típico 'se' ($|0.9862 - 0.3807| = 0.6055$). El primero aparece en las dos tablas, pero con valores más bien cercanos, por lo que su diferencia es relativamente

[19]De aquí en adelante se utilizan las cliticidades que no toman en cuenta la puntuación. Si bien este ejercicio no se puede escapar de la lengua escrita, no conviene depender tanto de ella, sobre todo si se quiere hacer generalizaciones sobre toda la lengua.

Tabla 3.3 (continuación):
Preformas gramaticales del *CEMC* en orden de *cliticidad*

	clítico	fr.	punt.	entrop.	econ.	cliticidad "general"	cliticidad "escrita"
67.	habían	443	0.9955	0.7136	0.07209	0.3928	0.5937
68.	estar	665	0.9624	0.7098	0.06794	0.3888	0.58
69.	otro	1979	0.8762	0.6874	0.0902	0.3888	0.5513
70.	haber	604	0.9785	0.7034	0.06854	0.386	0.5835
71.	eran	561	0.9643	0.7222	0.04959	0.3859	0.5787
72.	muchos	941	0.9362	0.6955	0.07598	0.3857	0.5692
73.	en	50123	0.9964	0.632	0.1384	0.3852	0.5889
74.	también	3501	0.8672	0.6978	0.07048	0.3842	0.5452
75.	bien	2603	0.7349	0.6618	0.1058	0.3838	0.5009
76.	sólo	1716	0.9924	0.714	0.05278	0.3834	0.5864
77.	tiene	3122	0.9622	0.6821	0.08344	0.3828	0.5759
78.	unos	1166	0.9811	0.71	0.05417	0.3821	0.5818
79.	algunos	822	0.9538	0.697	0.06577	0.3814	0.5722
80.	mejor	1110	0.8937	0.6956	0.06716	0.3814	0.5521
81.	era	2650	0.966	0.6932	0.06915	0.3812	0.5761
82.	ya	9791	0.9584	0.6807	0.08124	0.381	0.5735
83.	cualquier	655	0.9969	0.6987	0.05854	0.3786	0.5847
84.	donde	1968	0.9782	0.6861	0.07013	0.3781	0.5781
85.	hacer	1707	0.9192	0.6873	0.06871	0.378	0.5584
86.	estaban	402	0.9602	0.6966	0.05897	0.3778	0.5719
87.	estoy	693	0.9726	0.6893	0.06593	0.3776	0.5759
88.	hace	2003	0.9501	0.6891	0.06433	0.3767	0.5678
89.	diferentes	412	0.9053	0.6849	0.06547	0.3752	0.5519
90.	siempre	1655	0.9021	0.6924	0.05378	0.3731	0.5494
91.	hacen	733	0.9222	0.6794	0.0648	0.3721	0.5555
92.	sean	286	0.965	0.7048	0.03917	0.372	0.5697
93.	esas	658	0.9818	0.6825	0.06082	0.3717	0.575
94.	nuestros	445	0.973	0.677	0.06413	0.3706	0.5714
95.	yo	7044	0.9103	0.6694	0.07086	0.3702	0.5502
96.	bastante	510	0.8608	0.683	0.05713	0.37	0.5336
97.	pero	8336	0.9388	0.6805	0.05808	0.3693	0.5591
98.	nuestras	283	0.9647	0.6854	0.05305	0.3692	0.5677
99.	si	6122	0.9763	0.6684	0.06982	0.3691	0.5715
100.	tienen	1275	0.9545	0.6761	0.06126	0.3687	0.564

pequeña. De hecho, la diferencia negativa de valores garantiza que no pueda considerarse proclítico (no aparece en la tabla 3.5) y la poca diferencia hace sumamente cuestionable su carácter de enclítico (núm. 5 en tabla 3.6). Por otra parte, el segmento 'se', con una diferencia mayor de 0.6, es el proclítico más adherido a las formas que precede, según la tabla 3.5.

Estos ejemplos muestran que la medida que hemos llamado cliticidad es más bien una

Tabla 3.4: Postformas gramaticales del *CEMC* en orden de *cliticidad*

	clítico	fr.	punt.	entrop.	econ.	cliticidad "general"	cliticidad "escrita"
1.	y	60303	0.8072	0.903	1	0.9515	0.9034
2.	de	114346	0.9527	1	0.732	0.866	0.8949
3.	en	50123	0.8576	0.9357	0.7779	0.8568	0.8571
4.	a	45525	0.9037	0.9224	0.4697	0.6961	0.7653
5.	del	18621	0.9659	0.9592	0.4146	0.6869	0.7799
6.	con	18747	0.8347	0.8902	0.4148	0.6525	0.7132
7.	por	19835	0.7842	0.8262	0.3793	0.6028	0.6632
8.	para	14655	0.8373	0.8699	0.3171	0.5935	0.6747
9.	que	67243	0.8894	0.8166	0.3128	0.5647	0.673
10.	al	11179	0.8681	0.8783	0.2376	0.5579	0.6613
11.	o	8264	0.7861	0.817	0.2025	0.5097	0.6019
12.	el	51708	0.8783	0.6849	0.2029	0.4439	0.5887
13.	como	11088	0.7308	0.7112	0.1485	0.4299	0.5302
14.	sobre	2740	0.8504	0.7797	0.07685	0.4283	0.569
15.	entre	2528	0.8354	0.7672	0.07278	0.42	0.5585
16.	más	9778	0.936	0.7445	0.07047	0.4075	0.5837
17.	hacia	911	0.921	0.7555	0.03341	0.3945	0.57
18.	mucho	1754	0.9396	0.7295	0.03522	0.3823	0.5681
19.	se	33623	0.8517	0.6398	0.1215	0.3807	0.5377
20.	es	18601	0.8119	0.6667	0.09411	0.3804	0.5242
21.	hasta	3018	0.7969	0.7117	0.04649	0.3791	0.5184
22.	un	19765	0.927	0.6718	0.08137	0.3766	0.5601
23.	una	16473	0.9204	0.6723	0.0712	0.3718	0.5546
24.	también	3501	0.8326	0.7007	0.0387	0.3697	0.524
25.	durante	1017	0.8535	0.7146	0.02177	0.3682	0.53
26.	muy	4915	0.8844	0.6896	0.0449	0.3673	0.5397
27.	e	1325	0.8294	0.7104	0.02302	0.3667	0.5209
28.	son	4447	0.8451	0.6851	0.04436	0.3647	0.5248
29.	la	73110	0.9175	0.5912	0.1239	0.3575	0.5442
30.	tan	1591	0.8831	0.684	0.02475	0.3544	0.5306
31.	ante	621	0.8696	0.6891	0.01397	0.3515	0.5242
32.	nada	2730	0.8205	0.6505	0.02961	0.3401	0.5002
33.	desde	1904	0.7463	0.6498	0.02811	0.3389	0.4747

medida de la propiedad que tienen los segmentos de gramaticalizarse. una medida de "gramaticalidad" de los vocablos, no en el sentido de aceptabilidad intuitiva, sino en el de uso gramatical formal. La cliticidad propia es una forma de asociación direccional de los segmentos (ya sea hacia la derecha o a la izquierda) y, como se ve, se puede determinar comparando estos índices calculados en ambas direcciones[20].

[20]En este trabajo. en lugar de apresurarnos a darle un nuevo nombre al índice de cliticidad (por ej. "*gramaticalidad*", que en lingüística va significa varias cosas), seguiremos refiriéndonos a la cliticidad como a la fuerza de adhesión de un segmento a los vocablos de un corpus. por lo que de alguna manera nos estaremos

Tabla 3.4 (continuación):
Postformas gramaticales del *CEMC* en orden de *cliticidad*

	clítico	fr.	punt.	entrop.	econ.	cliticidad "general"	cliticidad "escrita"
34.	contra	675	0.9185	0.6598	0.01323	0.3365	0.5305
35.	cuando	4765	0.736	0.6326	0.03976	0.3362	0.4694
36.	así	4161	0.7186	0.6243	0.0416	0.3329	0.4615
37.	nacional	445	0.991	0.6431	0.02076	0.3319	0.5516
38.	aquí	4024	0.8497	0.633	0.02985	0.3314	0.5042
39.	algo	1109	0.9026	0.6421	0.01723	0.3297	0.5206
40.	conmigo	251	0.9602	0.6353	0.01816	0.3267	0.5379
41.	uno	3075	0.922	0.6303	0.02119	0.3258	0.5245
42.	sin	3179	0.6379	0.5982	0.05048	0.3243	0.4289
43.	bajo	627	0.8357	0.6355	0.0115	0.3235	0.4943
44.	ni	2392	0.6693	0.6162	0.03048	0.3233	0.4387
45.	antes	1288	0.8269	0.633	0.01185	0.3224	0.4906
46.	fue	2827	0.873	0.6166	0.02345	0.32	0.5043
47.	pa	788	0.7817	0.6125	0.0194	0.3159	0.4712
48.	las	20882	0.9299	0.5824	0.04636	0.3144	0.5196
49.	social	503	0.9742	0.6097	0.01796	0.3138	0.5339
50.	lo	13705	0.8772	0.5985	0.02644	0.3125	0.5007
51.	los	31231	0.915	0.5666	0.05755	0.3121	0.513
52.	totalmente	184	0.9402	0.6151	0.005975	0.3105	0.5204
53.	su	12520	0.9395	0.5828	0.03708	0.3099	0.5198
54.	mal	579	0.9845	0.6039	0.01467	0.3093	0.5343
55.	mexicana	265	0.9849	0.6009	0.01555	0.3082	0.5338
56.	libre	280	0.9536	0.6041	0.01222	0.3082	0.5233
57.	bastante	510	0.8608	0.6037	0.01212	0.3079	0.4922
58.	bien	2603	0.9355	0.5981	0.01722	0.3076	0.5169
59.	era	2650	0.8687	0.6003	0.01473	0.3075	0.4946
60.	fueron	861	0.8792	0.6003	0.01124	0.3058	0.4969
61.	allí	973	0.8798	0.5991	0.01176	0.3054	0.4969
62.	completamente	191	0.9267	0.6025	0.007662	0.3051	0.5123
63.	directamente	145	0.9379	0.6026	0.005237	0.3039	0.5153
64.	no	31676	0.6812	0.5393	0.06796	0.3036	0.4295
65.	casi	1133	0.7908	0.5937	0.01075	0.3022	0.4651
66.	ayer	434	0.894	0.5875	0.009966	0.2987	0.4971

Pero volviendo a las listas de clíticos, no está de más observar que los clíticos tradicionales del español aparecen otra vez concentrados al principio de la tabla (esta vez densamente apretados entre el núm. 1 y el 17 de la tabla 3.5), donde la diferencia entre cliticidades es la mayor. De todas maneras, se puede ver que la frontera entre clíticos y no clíticos no es clara (a partir del núm. 22 empiezan a ocurrir verbos finitos). pero es obvio que a menos

refiriendo, por ahora, tanto al fenómeno de adhesión de los propiamente clíticos. como al de los otros tipos de formas gramaticales mediante el mismo término sombrilla de 'cliticidad'.

Tabla 3.4 (continuación):
Postformas gramaticales del *CEMC* en orden de *cliticidad*

	clítico	fr.	punt.	entrop.	econ.	cliticidad "general"	cliticidad "escrita"
67.	están	1561	0.9052	0.5873	0.009755	0.2985	0.5008
68.	está	4108	0.8761	0.5792	0.01742	0.2983	0.4909
69.	perfectamente	136	0.9338	0.5887	0.007157	0.2979	0.5099
70.	siempre	1655	0.7994	0.5843	0.01066	0.2975	0.4648
71.	ahí	2060	0.8495	0.5774	0.01594	0.2967	0.481
72.	unos	1166	0.9014	0.5818	0.009892	0.2959	0.4977
73.	ya	9791	0.7108	0.553	0.03832	0.2957	0.434
74.	todavía	780	0.7974	0.5755	0.01191	0.2937	0.4616
75.	será	747	0.8153	0.5732	0.014	0.2936	0.4675
76.	todo	4119	0.8844	0.5754	0.0115	0.2934	0.4904
77.	después	2051	0.7928	0.5743	0.01075	0.2925	0.4593
78.	eran	561	0.8663	0.575	0.009315	0.2921	0.4835
79.	especial	311	0.9582	0.5748	0.007899	0.2914	0.5136
80.	si	6122	0.7117	0.5527	0.02915	0.2909	0.4312
81.	usted	1482	0.8772	0.5685	0.01201	0.2902	0.4859
82.	fuera	736	0.9484	0.5707	0.007532	0.2891	0.5089
83.	natural	261	0.9617	0.567	0.01075	0.2889	0.5131
84.	esto	1542	0.8346	0.57	0.007558	0.2888	0.4707
85.	yo	7044	0.7518	0.5468	0.03007	0.2884	0.4429
86.	total	483	0.9565	0.5661	0.01019	0.2882	0.5109
87.	debe	1318	0.8741	0.5628	0.01304	0.2879	0.4833
88.	demasiado	214	0.9252	0.5682	0.007478	0.2879	0.5003
89.	mexicano	391	0.9847	0.5642	0.009861	0.287	0.5196
90.	unas	643	0.8958	0.5651	0.007801	0.2864	0.4896
91.	muchas	851	0.8872	0.5649	0.007907	0.2864	0.4867
92.	internacional	159	1	0.5585	0.01417	0.2863	0.5242
93.	ahora	1923	0.792	0.561	0.01099	0.286	0.4547
94.	alguna	565	0.9469	0.5655	0.005533	0.2855	0.506
95.	tiene	3122	0.8824	0.555	0.01393	0.2845	0.4838
96.	común	262	0.9924	0.5593	0.009193	0.2843	0.5203
97.	mi	4840	0.875	0.5507	0.01727	0.284	0.481
98.	rápidamente	130	0.9308	0.5656	0.002121	0.2839	0.4995
99.	algunos	822	0.8917	0.5608	0.006883	0.2839	0.4865
100.	mayores	237	0.9705	0.557	0.0101	0.2836	0.5125

diferencia entre los valores de cliticidad, menos carácter de clítico del segmento examinado. Con respecto a los verbos, es de notarse que aquí (tabla 3.5) aparecen más formas verbales que en la tabla de preformas. Además de los verbos arriba enumerados que aquí aparecen representados por más segmentos, en la tabla 3.5 ocurren otros que corresponden a 'dar', 'quedar', 'querer' y 'decir' (núms. 66, 72, 84, 90, y 91). Si bien no podemos considerar a estos últimos como clíticos verdaderos, es interesante que casi todos estos verbos sirvan en

Tabla 3.5: Proclíticos del *CEMC* en orden de *cliticidad*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
1.	se	33623	0.3807	0.9862	0.6055	0.5972
2.	la	73110	0.3575	0.9243	0.5668	0.5239
3.	los	31231	0.3121	0.8691	0.557	0.4841
4.	su	12520	0.3099	0.8538	0.5439	0.4643
5.	un	19765	0.3766	0.8745	0.4979	0.4354
6.	una	16473	0.3718	0.8362	0.4645	0.3884
7.	el	51708	0.4439	0.8595	0.4156	0.3572
8.	las	20882	0.3144	0.732	0.4176	0.3057
9.	me	10410	0.2709	0.6888	0.4179	0.2878
10.	sus	5267	0.2826	0.6948	0.4122	0.2864
11.	le	8502	0.2552	0.617	0.3618	0.2232
12.	nos	3399	0.2711	0.5977	0.3266	0.1952
13.	te	3389	0.2425	0.5413	0.2988	0.1617
14.	ser	2830	0.2356	0.5262	0.2906	0.1529
15.	les	2021	0.2368	0.5156	0.2788	0.1438
16.	esta	2925	0.2673	0.5261	0.2588	0.1361
17.	lo	13705	0.3125	0.5455	0.233	0.1271
18.	más	9778	0.4075	0.6127	0.2052	0.1258
19.	dos	3216	0.282	0.5186	0.2366	0.1227
20.	muy	4915	0.3673	0.5758	0.2085	0.1201
21.	esa	1666	0.2703	0.4936	0.2233	0.1102
22.	había	2023	0.2537	0.4733	0.2196	0.1039
23.	al	11179	0.5579	0.6966	0.1387	0.09664
24.	ese	1977	0.2666	0.4693	0.2027	0.09511
25.	está	4108	0.2983	0.4914	0.1931	0.09486
26.	puede	2480	0.2679	0.4671	0.1992	0.09303
27.	otros	1499	0.2211	0.434	0.2129	0.09238
28.	tan	1591	0.3544	0.5245	0.1701	0.08919
29.	mi	4840	0.284	0.4695	0.1855	0.08709
30.	están	1561	0.2985	0.477	0.1785	0.08515
31.	han	1947	0.2431	0.4333	0.1902	0.0824
32.	cuya	267	0.2057	0.405	0.1992	0.08069
33.	tu	1278	0.2682	0.448	0.1798	0.08054

la construcción de perífrasis verbales de los tipos más variados (como si con unos cuantos verbos muy frecuentes se construyeran muchísimos sintagmas verbales con los otros verbos. los de contenido que son muchísimos más, pero mucho menos frecuentes). De hecho, no es extraño que un verbo (con todo y sus formas flexionadas) se adhiera a palabras plenas para formar nuevos paradigmas verbales (por ej., en español, algunas formas del verbo 'haber' se pegaron a los infinitivos para formar los paradigmas verbales del futuro y del copretérito o

Tabla 3.5 (continuación):
Proclíticos del *CEMC* en orden de *cliticidad*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
34.	esos	741	0.2381	0.4264	0.1884	0.08033
35.	ha	4105	0.2527	0.4363	0.1836	0.08008
36.	estas	765	0.2605	0.4355	0.175	0.07621
37.	estos	1005	0.2477	0.4253	0.1776	0.07552
38.	nuestra	768	0.2294	0.4115	0.1821	0.07493
39.	no	31676	0.3036	0.4601	0.1565	0.07201
40.	haber	604	0.2017	0.386	0.1843	0.07112
41.	otras	1071	0.2227	0.3997	0.177	0.07074
42.	que	67243	0.5647	0.6667	0.102	0.06797
43.	hay	4013	0.2451	0.4087	0.1637	0.0669
44.	habían	443	0.2286	0.3928	0.1642	0.06451
45.	pueden	944	0.2727	0.4213	0.1486	0.06262
46.	sin	3179	0.3243	0.4595	0.1351	0.06208
47.	hacen	733	0.2056	0.3721	0.1665	0.06196
48.	grandes	845	0.2582	0.4089	0.1507	0.06164
49.	son	4447	0.3647	0.4872	0.1225	0.0597
50.	nuestros	445	0.2095	0.3706	0.161	0.05968
51.	tres	1598	0.2662	0.409	0.1427	0.05836
52.	estoy	693	0.2243	0.3776	0.1533	0.0579
53.	nuestras	283	0.2143	0.3692	0.1549	0.05718
54.	pudo	235	0.2019	0.3589	0.1571	0.05638
55.	cualquier	655	0.2311	0.3786	0.1475	0.05586
56.	mis	963	0.2684	0.4059	0.1375	0.05582
57.	estaba	1368	0.2666	0.4045	0.1379	0.05578
58.	estar	665	0.2457	0.3888	0.1431	0.05565
59.	podemos	380	0.2211	0.3677	0.1466	0.0539
60.	pero	8336	0.2235	0.3693	0.1458	0.05384
61.	menos	1413	0.2234	0.367	0.1437	0.05274
62.	esas	658	0.2311	0.3717	0.1406	0.05226
63.	hizo	737	0.2065	0.3538	0.1474	0.05214
64.	otro	1979	0.2552	0.3888	0.1335	0.05192
65.	hacer	1707	0.2417	0.378	0.1363	0.05151
66.	dar	755	0.2039	0.3506	0.1467	0.05144

condicional[21]).

El caso es que en la tabla 3.5 de proclíticos, todas estas formas verbales (cuyo carácter de clítico es cuestionable), aparecen mezcladas con otros segmentos que sí gozan de cierto carácter clitical: adjetivos posesivos ('mi', 'tu', 'mis') y demostrativos ('esa', 'esos', 'estos',

[21] Véanse Company, "Los futuros en el español medieval. Sus orígenes y evolución" [42]. *Nueva Revista de Filología Española*, 34 (1985-86), pp. 48-107 y, para una posible explicación de por qué convivieron tantos siglos los futuros analíticos con los sintéticos, Company y Medina, "Sintaxis motivada pragmáticamente. Futuros analíticos y futuros sintéticos" [41], *Revista de Filología Española*, LXXIX. 1999, pp. 65-100.

Tabla 3.5 (continuación):
Proclíticos del *CEMC* en orden de *cliticidad*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
67.	nuestro	802	0.2101	0.355	0.1449	0.05143
68.	aquel	398	0.2145	0.3539	0.1394	0.04934
69.	hace	2003	0.2461	0.3767	0.1306	0.04921
70.	diferentes	412	0.2453	0.3752	0.1299	0.04873
71.	buena	519	0.2199	0.3557	0.1357	0.04828
72.	quedó	309	0.2031	0.3433	0.1402	0.04812
73.	sean	286	0.2437	0.372	0.1283	0.04774
74.	solo	405	0.2228	0.3549	0.132	0.04686
75.	muchos	941	0.2649	0.3857	0.1208	0.04661
76.	debe	1318	0.2879	0.4032	0.1153	0.04649
77.	hemos	523	0.2238	0.3548	0.131	0.04646
78.	esté	275	0.2206	0.3519	0.1312	0.04618
79.	siendo	355	0.211	0.3447	0.1337	0.0461
80.	estaban	402	0.2566	0.3778	0.1212	0.0458
81.	aquella	340	0.2252	0.353	0.1278	0.04512
82.	de	114346	0.866	0.9148	0.04879	0.04463
83.	podría	367	0.2491	0.3683	0.1191	0.04387
84.	dicha	176	0.2139	0.3414	0.1275	0.04351
85.	sea	1608	0.2116	0.339	0.1273	0.04315
86.	unas	643	0.2864	0.3955	0.1091	0.04315
87.	mayor	1209	0.251	0.368	0.117	0.04305
88.	estamos	475	0.2241	0.3478	0.1237	0.04302
89.	donde	1968	0.2644	0.3781	0.1137	0.04299
90.	queda	405	0.2336	0.3534	0.1197	0.04231
91.	quería	345	0.2045	0.3317	0.1272	0.04221
92.	mejor	1110	0.2714	0.3814	0.1099	0.04192
93.	tengo	1126	0.2211	0.3428	0.1217	0.04173
94.	diversas	190	0.2012	0.3278	0.1266	0.0415
95.	diversos	229	0.2022	0.3283	0.1261	0.0414
96.	tienen	1275	0.2578	0.3687	0.1109	0.04087
97.	pues	6077	0.2079	0.3309	0.123	0.04071
98.	luego	1957	0.2097	0.332	0.1223	0.0406
99.	hubiera	362	0.2151	0.3356	0.1205	0.04043
100.	cierta	184	0.2269	0.3439	0.117	0.04025

'estas', etc.), adverbios monosílabos ('más', 'muy' y 'tan'). No hay que olvidar que otros segmentos cualitativamente similares ocurrieron al principio de la tabla mezclados con los pronombres proclíticos ('su', 'sus', 'un', 'una', etc.).

En cuanto a la tabla 3.6, nótese que casi todos los "enclíticos" obtenidos del corpus tienen una diferencia muy pequeña entre cliticidades de cada lado[22]. Destaca el hecho de que las

[22]Recuérdese que los pronombres enclíticos verdaderos del español aparecen sufijados en las palabras gráficas, por lo que fueron tratados como afijos en el capítulo anterior. Aunque en ciertos contextos la

Tabla 3.6: "Enclíticos" del *CEMC* en orden de *cliticidad*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
1.	en	50123	0.3852	0.8568	0.4716	0.404
2.	a	45525	0.4763	0.6961	0.2198	0.153
3.	con	18747	0.4609	0.6525	0.1916	0.125
4.	por	19835	0.4012	0.6028	0.2016	0.1215
5.	y	60303	0.8478	0.9515	0.1037	0.09864
6.	sobre	2740	0.2413	0.4283	0.1869	0.08005
7.	hacia	911	0.2215	0.3945	0.173	0.06823
8.	entre	2528	0.2846	0.42	0.1353	0.05684
9.	durante	1017	0.2142	0.3682	0.154	0.0567
10.	ante	621	0.2038	0.3515	0.1478	0.05194
11.	conmigo	251	0.2067	0.3267	0.1201	0.03923
12.	contra	675	0.229	0.3365	0.1075	0.03619
13.	nada	2730	0.2346	0.3401	0.1055	0.03586
14.	antes	1288	0.2214	0.3224	0.101	0.03257
15.	desde	1904	0.253	0.3389	0.08595	0.02913
16.	social	503	0.226	0.3138	0.08782	0.02756
17.	nacional	445	0.2513	0.3319	0.08066	0.02678
18.	bajo	627	0.245	0.3235	0.07855	0.02541
19.	hasta	3018	0.314	0.3791	0.06515	0.0247
20.	natural	261	0.2049	0.2889	0.08397	0.02425
21.	libre	280	0.2297	0.3082	0.07843	0.02417
22.	mexicana	265	0.2306	0.3082	0.07762	0.02393
23.	mucho	1754	0.3214	0.3823	0.06092	0.02329
24.	total	483	0.2082	0.2882	0.08001	0.02305
25.	internacional	159	0.2069	0.2863	0.07937	0.02272
26.	después	2051	0.2224	0.2925	0.07014	0.02052
27.	juntos	103	0.2016	0.2752	0.07367	0.02027
28.	del	18621	0.6578	0.6869	0.02906	0.01996
29.	real	181	0.2098	0.2785	0.06867	0.01912
30.	para	14655	0.5623	0.5935	0.03119	0.01851
31.	normal	182	0.2117	0.278	0.06631	0.01844
32.	directamente	145	0.2436	0.3039	0.06032	0.01833
33.	así	4161	0.2831	0.3329	0.04987	0.0166

formas con mayor diferencia son casi todas preposiciones. Así, rodeada de éstas en la tabla, aparece en el núm. 5 la conjunción 'y', que en realidad no exhibe una gran diferencia entre cliticidades (apenas de 0.1037). Al examinar el resto de la lista, vemos que predominan formas adjetivas ('social', 'nacional', 'bajo' —que también es preposición—, 'natural', 'libre',

tradición los sufije a las palabras gráficas, cabe presumir que las cantidades de afijalidad/cliticidad de los enclíticos serán parecidas a los valores de los proclíticos, si es que son tipos de clíticos comparables. Naturalmente, para corroborar esto será necesario construir una estructura que proporcione los datos adecuados; por ej. una estructura arbórea que incluya, además de los caracteres al interior de los vocablos las distribuciones de vocablos anteriores y posteriores.

Tabla 3.6 (continuación):
"Enclíticos" del *CEMC* en orden de *cliticidad*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
34.	común	262	0.2273	0.2843	0.05701	0.01621
35.	inicial	90	0.2092	0.2685	0.05927	0.01591
36.	militar	87	0.2146	0.2707	0.05611	0.01519
37.	cultural	149	0.2139	0.27	0.05616	0.01517
38.	anteriores	156	0.2201	0.2742	0.05402	0.01481
39.	comercial	111	0.2192	0.2731	0.05391	0.01472
40.	todos	2627	0.2167	0.2707	0.05402	0.01462
41.	semejante	105	0.2004	0.2568	0.05639	0.01448
42.	humana	168	0.2189	0.2718	0.05291	0.01438
43.	superior	267	0.2199	0.2715	0.05165	0.01402
44.	cincuenta	305	0.2113	0.2631	0.05175	0.01361
45.	adecuada	126	0.2087	0.2607	0.05202	0.01356
46.	industrial	167	0.222	0.2716	0.04962	0.01348
47.	comerciales	98	0.2023	0.2548	0.05255	0.01339
48.	político	187	0.2193	0.2689	0.04961	0.01334
49.	sí	9079	0.2306	0.2782	0.04761	0.01324
50.	popular	136	0.2302	0.2776	0.04745	0.01317
51.	ayer	434	0.2547	0.2987	0.04397	0.01314
52.	especial	311	0.2465	0.2914	0.04486	0.01307
53.	blanca	144	0.2095	0.2597	0.05024	0.01305
54.	original	112	0.2203	0.2682	0.04784	0.01283
55.	exclusivamente	96	0.2156	0.2634	0.04782	0.0126
56.	local	124	0.2118	0.26	0.04823	0.01254
57.	mundial	149	0.2164	0.2638	0.0474	0.0125
58.	general	850	0.2062	0.2552	0.04897	0.0125
59.	atrás	224	0.2064	0.2538	0.04738	0.01202
60.	algo	1109	0.2933	0.3297	0.0364	0.012
61.	legal	61	0.2017	0.2489	0.04724	0.01176
62.	sociales	194	0.2099	0.2557	0.04584	0.01172
63.	mediante	310	0.2285	0.2708	0.04234	0.01147
64.	profesional	113	0.2257	0.2678	0.04211	0.01128
65.	inferior	135	0.2095	0.2538	0.04427	0.01124
66.	suficiente	210	0.2099	0.254	0.04405	0.01119

'mexicana', 'total', 'internacional', 'juntos', 'real', 'normal', 'común', 'inicial', 'militar', 'cultural', 'anteriores', 'comercial', 'semejante', 'humana', etc.), cosa que no sorprende mucho al tratarse de segmentos de ocurrencia profusa que se adhieren a los sustantivos que siguen. Las primeras y poquísimas formas verbales representan a los verbos 'llegar', 'ir' y 'hacer' y aparecen casi al final de la tabla (núms. 84, 90, 91 y 100). De esta manera, por inspección podemos constatar que no se trata, tampoco cualitativamente, de ningún segmento que la tradición

Tabla 3.6 (continuación):
"Enclíticos" del *CEMC* en orden de *cliticidad*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
67.	oficiales	73	0.2022	0.247	0.04477	0.01106
68.	uno	3075	0.2926	0.3258	0.03314	0.0108
69.	completa	141	0.239	0.2777	0.0387	0.01075
70.	toda	1191	0.2343	0.2731	0.03884	0.01061
71.	igual	505	0.2231	0.2634	0.04026	0.0106
72.	naturales	147	0.2098	0.2518	0.04203	0.01058
73.	artística	72	0.2063	0.2488	0.04249	0.01057
74.	claramente	99	0.2384	0.2764	0.03803	0.01051
75.	rápidamente	130	0.2474	0.2839	0.03651	0.01036
76.	lentamente	82	0.2241	0.2634	0.03921	0.01033
77.	física	101	0.2055	0.2466	0.04111	0.01014
78.	aproximadamente	198	0.2245	0.2622	0.03768	0.009879
79.	permanente	84	0.2185	0.2567	0.03828	0.009827
80.	o	8264	0.4912	0.5097	0.01856	0.009463
81.	abierta	77	0.2099	0.248	0.03811	0.00945
82.	mexicano	391	0.2544	0.287	0.0326	0.009358
83.	iguales	98	0.2043	0.2428	0.03849	0.009344
84.	llegó	400	0.2074	0.2443	0.03696	0.009029
85.	inmediata	53	0.2095	0.2457	0.03628	0.008914
86.	blanco	252	0.2096	0.2453	0.03569	0.008754
87.	aquí	4024	0.3057	0.3314	0.02576	0.008537
88.	económica	240	0.2346	0.2666	0.03194	0.008513
89.	mexicanas	73	0.2078	0.2422	0.03447	0.008349
90.	iba	677	0.2099	0.2431	0.03329	0.008093
91.	llega	358	0.2164	0.2488	0.03239	0.008057
92.	personal	289	0.2407	0.2685	0.02781	0.007468
93.	preparado	56	0.2038	0.2343	0.03054	0.007157
94.	perfecto	67	0.2098	0.2397	0.02983	0.007148
95.	vivo	128	0.2327	0.2599	0.02718	0.007064
96.	abierto	99	0.232	0.259	0.02701	0.006994
97.	nuevamente	112	0.2492	0.2743	0.02515	0.0069
98.	viva	114	0.2019	0.2303	0.02841	0.006543
99.	seguro	208	0.2148	0.2415	0.02671	0.006451
100.	hecha	109	0.2127	0.2396	0.02692	0.006449

considere un enclítico verdadero[23]. Si bien los datos cuantitativos parecen indicadores necesarios para examinar la naturaleza de los enclíticos del español, la información disponible en esta investigación es interesante pero no suficiente para llegar a alguna conclusión sobre dicha naturaleza.

[23]De todos modos, valdría la pena examinar el hecho de que tantas preposiciones se concentren con los valores más altos al principio de la tabla. También sería interesante analizar el pronombre 'conmigo' (núm. 11) que, sin ser clítico, tiende a ocurrir después de algo.

De todo esto se puede observar que, al hablar de la cliticidad de un segmento, es necesario especificar su dirección de adhesión, es decir, hablar de procliticidad y encliticidad. De esta manera, para cada segmento, se trate o no de un clítico, hay dos valores de cliticidad (uno como preforma y otro como postforma). Estos valores, sin embargo, así como se pueden restar para determinar si se trata de proclíticos o enclíticos, también se pueden sumar para obtener un índice general de lo que podemos llamar asociación gramatical con respecto al resto de los segmentos del corpus.

Tabla 3.7: Formas gramaticales del *CEMC* en orden de *cliticidad* total

			cliticidad			cliticidad
	clítico	fr.	de un lado	del otro	diferencia	total
1.	y	60303	0.9515	0.8478	-0.1037	1.799
2.	de	114346	0.866	0.9148	0.04879	1.781
3.	se	33623	0.3807	0.9862	0.6055	1.367
4.	del	18621	0.6869	0.6578	-0.02906	1.345
5.	el	51708	0.4439	0.8595	0.4156	1.303
6.	la	73110	0.3575	0.9243	0.5668	1.282
7.	al	11179	0.5579	0.6966	0.1387	1.255
8.	un	19765	0.3766	0.8745	0.4979	1.251
9.	en	50123	0.8568	0.3852	-0.4716	1.242
10.	que	67243	0.5647	0.6667	0.102	1.231
11.	una	16473	0.3718	0.8362	0.4645	1.208
12.	los	31231	0.3121	0.8691	0.557	1.181
13.	a	45525	0.6961	0.4763	-0.2198	1.172
14.	su	12520	0.3099	0.8538	0.5439	1.164
15.	para	14655	0.5935	0.5623	-0.03119	1.156
16.	con	18747	0.6525	0.4609	-0.1916	1.113
17.	las	20882	0.3144	0.732	0.4176	1.046
18.	más	9778	0.4075	0.6127	0.2052	1.02
19.	por	19835	0.6028	0.4012	-0.2016	1.004
20.	o	8264	0.5097	0.4912	-0.01856	1.001
21.	sus	5267	0.2826	0.6948	0.4122	0.9774
22.	me	10410	0.2709	0.6888	0.4179	0.9597
23.	muy	4915	0.3673	0.5758	0.2085	0.9431
24.	como	11088	0.4299	0.4789	0.04902	0.9088
25.	tan	1591	0.3544	0.5245	0.1701	0.8788
26.	le	8502	0.2552	0.617	0.3618	0.8722
27.	nos	3399	0.2711	0.5977	0.3266	0.8688
28.	lo	13705	0.3125	0.5455	0.233	0.858
29.	son	4447	0.3647	0.4872	0.1225	0.852
30.	es	18601	0.3804	0.4359	0.05555	0.8163
31.	dos	3216	0.282	0.5186	0.2366	0.8006
32.	esta	2925	0.2673	0.5261	0.2588	0.7934
33.	está	4108	0.2983	0.4914	0.1931	0.7897

Tabla 3.7 (continuación):
Formas gramaticales del *CEMC* en orden de *cliticidad* total

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	cliticidad total
34.	te	3389	0.2425	0.5413	0.2988	0.7838
35.	sin	3179	0.3243	0.4595	0.1351	0.7838
36.	e	1325	0.3667	0.4122	0.04549	0.7789
37.	están	1561	0.2985	0.477	0.1785	0.7756
38.	esa	1666	0.2703	0.4936	0.2233	0.7639
39.	no	31676	0.3036	0.4601	0.1565	0.7637
40.	ser	2830	0.2356	0.5262	0.2906	0.7618
41.	también	3501	0.3697	0.3842	0.01443	0.7539
42.	mi	4840	0.284	0.4695	0.1855	0.7535
43.	les	2021	0.2368	0.5156	0.2788	0.7525
44.	ese	1977	0.2666	0.4693	0.2027	0.7358
45.	cuando	4765	0.3362	0.3995	0.06335	0.7357
46.	puede	2480	0.2679	0.4671	0.1992	0.735
47.	fue	2827	0.32	0.4097	0.08972	0.7297
48.	había	2023	0.2537	0.4733	0.2196	0.727
49.	ni	2392	0.3233	0.3942	0.07091	0.7176
50.	tu	1278	0.2682	0.448	0.1798	0.7161
51.	fueron	861	0.3058	0.403	0.09718	0.7087
52.	entre	2528	0.42	0.2846	-0.1353	0.7046
53.	mucho	1754	0.3823	0.3214	-0.06092	0.7038
54.	estas	765	0.2605	0.4355	0.175	0.696
55.	pueden	944	0.2727	0.4213	0.1486	0.694
56.	hasta	3018	0.3791	0.314	-0.06515	0.6931
57.	bien	2603	0.3076	0.3838	0.07617	0.6915
58.	debe	1318	0.2879	0.4032	0.1153	0.6911
59.	ha	4105	0.2527	0.4363	0.1836	0.6889
60.	era	2650	0.3075	0.3812	0.07368	0.6887
61.	unas	643	0.2864	0.3955	0.1091	0.682
62.	eran	561	0.2921	0.3859	0.09376	0.678
63.	bastante	510	0.3079	0.37	0.06212	0.678
64.	unos	1166	0.2959	0.3821	0.08623	0.6779
65.	ya	9791	0.2957	0.381	0.08529	0.6766
66.	han	1947	0.2431	0.4333	0.1902	0.6764

Es en este sentido que los segmentos que aparecen en la tabla 3.7 son los segmentos "más gramaticales" del *CEMC*. Están ordenados según el índice de ordenamiento (última columna) que aquí es sencillamente la suma de cliticidades relativas a los lados de cada segmento. Nótese que sin haber tomado en cuenta la longitud de los segmentos como factor de cliticidad, en esta tabla la mayoría de las formas son muy cortas (sobre todo entre las primeras 50). Nótese también que los pronombres proclíticos aparecen más repartidos que en la tablas anteriores (están entre el 3 y el 43) y, aunque no tan concentrados entre los primeros

Tabla 3.7 (continuación):
Formas gramaticales del *CEMC* en orden de *cliticidad* total

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	cliticidad total
67.	tres	1598	0.2662	0.409	0.1427	0.6752
68.	mis	963	0.2684	0.4059	0.1375	0.6743
69.	estos	1005	0.2477	0.4253	0.1776	0.673
70.	estaba	1368	0.2666	0.4045	0.1379	0.6712
71.	siempre	1655	0.2975	0.3731	0.07561	0.6706
72.	sobre	2740	0.4283	0.2413	-0.1869	0.6696
73.	tiene	3122	0.2845	0.3828	0.09828	0.6672
74.	grandes	845	0.2582	0.4089	0.1507	0.6671
75.	algunos	822	0.2839	0.3814	0.09755	0.6653
76.	esos	741	0.2381	0.4264	0.1884	0.6645
77.	sólo	1716	0.279	0.3834	0.1044	0.6624
78.	si	6122	0.2909	0.3691	0.0782	0.6601
79.	yo	7044	0.2884	0.3702	0.08174	0.6586
80.	otros	1499	0.2211	0.434	0.2129	0.6551
81.	hay	4013	0.2451	0.4087	0.1637	0.6538
82.	mejor	1110	0.2714	0.3814	0.1099	0.6528
83.	muchos	941	0.2649	0.3857	0.1208	0.6507
84.	casi	1133	0.3022	0.3457	0.04346	0.6479
85.	mal	579	0.3093	0.3383	0.02905	0.6476
86.	será	747	0.2936	0.3517	0.05811	0.6453
87.	otro	1979	0.2552	0.3888	0.1335	0.644
88.	donde	1968	0.2644	0.3781	0.1137	0.6425
89.	porque	5087	0.2781	0.3641	0.086	0.6423
90.	nuestra	768	0.2294	0.4115	0.1821	0.6409
91.	demasiado	214	0.2879	0.351	0.06316	0.6389
92.	allí	973	0.3054	0.3319	0.02641	0.6373
93.	aquí	4024	0.3314	0.3057	-0.02576	0.6371
94.	estar	665	0.2457	0.3888	0.1431	0.6346
95.	estaban	402	0.2566	0.3778	0.1212	0.6344
96.	completamente	191	0.3051	0.3267	0.0216	0.6317
97.	ahí	2060	0.2967	0.3329	0.0362	0.6295
98.	varios	438	0.2618	0.3675	0.1057	0.6293
99.	usted	1482	0.2902	0.3378	0.04753	0.628
100.	algunas	561	0.2808	0.3468	0.06605	0.6276

como en las otras tablas, se encuentran de todas maneras entre las formas más gramaticales. Con excepción de las formas verbales y algunos adverbios y adjetivos (concentrados sobre todo en la segunda parte de la tabla), la tabla es una lista de partes invariables de la oración. Las variables son formas de los verbos 'ser', 'estar', 'haber', 'poder', 'deber' y 'tener' (imprescindibles para cualquier descripción gramatical del español); los adverbios también son los básicos y necesarios (incluidos los monosilábicos): 'más', 'muy', 'tan', 'ya', 'bien', 'también',

'mucho', 'bastante', 'siempre', etc; similarmente, los pocos adjetivos son de uso profuso y ocurren antes de sustantivos ('grandes', 'mejor', etc.; hay incluso cuantificadores: 'dos' y 'tres'). Por último, en esta tabla aparecen más conjunciones subordinadoras ('que', 'como', 'cuando', 'donde' y 'porque') que en todas las tablas de esta sección.

Tabla 3.8: "Nexos" del *CEMC*

			cliticidad			índice de
	clítico	fr.	de un lado	del otro	diferencia	ordenamiento
1.	de	114346	0.866	0.9148	0.04879	0.8198
2.	y	60303	0.9515	0.8478	-0.1037	0.7555
3.	del	18621	0.6869	0.6578	-0.02906	0.63
4.	para	14655	0.5935	0.5623	-0.03119	0.5327
5.	que	67243	0.5647	0.6667	0.102	0.4784
6.	o	8264	0.5097	0.4912	-0.01856	0.4733
7.	al	11179	0.5579	0.6966	0.1387	0.4468
8.	como	11088	0.4299	0.4789	0.04902	0.3859
9.	también	3501	0.3697	0.3842	0.01443	0.3558
10.	es	18601	0.3804	0.4359	0.05555	0.3319
11.	e	1325	0.3667	0.4122	0.04549	0.3262
12.	a	45525	0.6961	0.4763	-0.2198	0.3259
13.	con	18747	0.6525	0.4609	-0.1916	0.3255
14.	totalmente	184	0.3105	0.307	-0.003501	0.3036
15.	todo	4119	0.2934	0.2911	-0.002357	0.2887
16.	esto	1542	0.2888	0.2888	0.00007197	0.2887
17.	completamente	191	0.3051	0.3267	0.0216	0.2849
18.	cuando	4765	0.3362	0.3995	0.06335	0.2829
19.	mal	579	0.3093	0.3383	0.02905	0.2827
20.	aquí	4024	0.3314	0.3057	-0.02576	0.2819
21.	allí	973	0.3054	0.3319	0.02641	0.2811
22.	pa	788	0.3159	0.298	-0.01796	0.281
23.	todavía	780	0.2937	0.3087	0.01501	0.2795
24.	allá	1229	0.2819	0.2874	0.005443	0.2766
25.	mayores	237	0.2836	0.2911	0.00755	0.2762
26.	perfectamente	136	0.2979	0.2867	-0.01128	0.2758
27.	muchas	851	0.2864	0.2981	0.01166	0.2752
28.	son	4447	0.3647	0.4872	0.1225	0.273
29.	más	9778	0.4075	0.6127	0.2052	0.271
30.	mucho	1754	0.3823	0.3214	-0.06092	0.2702
31.	por	19835	0.6028	0.4012	-0.2016	0.267
32.	ni	2392	0.3233	0.3942	0.07091	0.2652
33.	únicamente	216	0.2788	0.2935	0.01477	0.2647
34.	ahí	2060	0.2967	0.3329	0.0362	0.2644

Otro conjunto de segmentos interesante es aquel de las formas con cliticidades equivalentes, es decir, con tanta cliticidad de un lado como del otro. Intuitivamente, estas formas

Tabla 3.8 (continuación):
"Nexos" del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
35.	casi	1133	0.3022	0.3457	0.04346	0.2642
36.	ligeramente	80	0.2678	0.2719	0.004143	0.2637
37.	uno	3075	0.3258	0.2926	-0.03314	0.2628
38.	hoy	743	0.2778	0.295	0.01719	0.2616
39.	fácilmente	137	0.2816	0.2712	-0.01047	0.2611
40.	algo	1109	0.3297	0.2933	-0.0364	0.2609
41.	hasta	3018	0.3791	0.314	-0.06515	0.26
42.	feliz	162	0.2586	0.2601	0.001491	0.2571
43.	bastante	510	0.3079	0.37	0.06212	0.2562
44.	ahora	1923	0.286	0.3212	0.03517	0.2547
45.	usté	728	0.278	0.3061	0.02808	0.2525
46.	principal	258	0.2675	0.2862	0.01872	0.25
47.	fue	2827	0.32	0.4097	0.08972	0.2499
48.	plenamente	53	0.252	0.2541	0.002109	0.2499
49.	usted	1482	0.2902	0.3378	0.04753	0.2494
50.	entonces	2974	0.2749	0.303	0.02818	0.2493
51.	era	2650	0.3075	0.3812	0.07368	0.2481
52.	fuera	736	0.2891	0.2678	-0.02136	0.248
53.	actual	285	0.2497	0.2518	0.002062	0.2477
54.	mexicanos	339	0.2523	0.2576	0.005297	0.2471
55.	bien	2603	0.3076	0.3838	0.07617	0.2466
56.	acá	547	0.2488	0.252	0.003247	0.2456
57.	será	747	0.2936	0.3517	0.05811	0.2451
58.	tanto	1418	0.2727	0.3049	0.0322	0.2439
59.	orita	533	0.263	0.2847	0.02167	0.243
60.	viendo	167	0.2469	0.2512	0.004267	0.2427
61.	conocido	135	0.2533	0.2475	-0.005891	0.2417
62.	veinte	340	0.2473	0.2444	-0.002876	0.2416
63.	alguna	565	0.2855	0.3384	0.05288	0.2409
64.	así	4161	0.3329	0.2831	-0.04987	0.2407
65.	trabajando	244	0.2521	0.2458	-0.006288	0.2397
66.	tan	1591	0.3544	0.5245	0.1701	0.2395
67.	diferente	175	0.2403	0.2411	0.0008496	0.2394
68.	saber	431	0.2493	0.2597	0.01034	0.2394

deberían ser las más usadas como nexos. En la tabla 3.8, titulada "Nexos" del *CEMC*", están los segmentos cuyos valores de cliticidad, además de ser los más altos, difieren menos entre sí. Como en las tablas comentadas arriba, después de las columnas de cliticidad, aparece su diferencia en la columna siguiente. Al igual que en la tabla anterior, ésta puede ser positiva o negativa. La última columna contiene un índice de ordenamiento calculado a partir del producto de la menor de las cliticidades por el cociente de la menor sobre la mayor (si

Tabla 3.8 (continuación):
"Nexos" del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
69.	u	149	0.2746	0.2563	-0.01828	0.2392
70.	inmediatamente	151	0.2798	0.2582	-0.0216	0.2382
71.	otra	2082	0.2778	0.3249	0.04715	0.2374
72.	siempre	1655	0.2975	0.3731	0.07561	0.2372
73.	muerto	187	0.2442	0.2404	-0.003794	0.2367
74.	viene	555	0.2625	0.2914	0.02895	0.2364
75.	vino	353	0.2378	0.2394	0.001638	0.2361
76.	demasiado	214	0.2879	0.351	0.06316	0.2361
77.	ir	824	0.2543	0.2446	-0.009684	0.2353
78.	99[a]	16449	0.2751	0.3224	0.04724	0.2348
79.	muy	4915	0.3673	0.5758	0.2085	0.2343
80.	debidamente	67	0.2546	0.2768	0.02225	0.2341
81.	diez	554	0.249	0.266	0.017	0.2331
82.	cien	166	0.2375	0.2351	-0.002355	0.2328
83.	fueron	861	0.3058	0.403	0.09718	0.232
84.	fuerte	257	0.2414	0.2513	0.009874	0.2319
85.	propiamente	89	0.2322	0.2314	-0.0008128	0.2306
86.	ya	9791	0.2957	0.381	0.08529	0.2295
87.	aún	489	0.2626	0.3007	0.03807	0.2294
88.	vienen	352	0.2514	0.2756	0.0242	0.2294
89.	si	6122	0.2909	0.3691	0.0782	0.2293
90.	el	51708	0.4439	0.8595	0.4156	0.2293
91.	unos	1166	0.2959	0.3821	0.08623	0.2291
92.	sin	3179	0.3243	0.4595	0.1351	0.229
93.	dinero	615	0.2572	0.2423	-0.01489	0.2283
94.	posteriormente	147	0.2289	0.2281	-0.0007254	0.2274
95.	particular	193	0.2282	0.2291	0.0008828	0.2274
96.	algunas	561	0.2808	0.3468	0.06605	0.2273
97.	va	1907	0.244	0.2621	0.01811	0.2271
98.	buscando	106	0.2402	0.2543	0.01408	0.2269
99.	dura	94	0.2344	0.2423	0.007973	0.2267
100.	constante	161	0.2711	0.2478	-0.02337	0.2264

[a]'99' representa cualquier combinación de dígitos.

la diferencia es poca, el cociente es casi igual a la unidad): $min(CL(s_i)) \times \frac{min(CL(s_i))}{max(CL(s_i))}$ (esto garantiza que no aparezcan en esta tabla los segmentos de baja cliticidad en ambos lados). De esta manera, además de las conjunciones 'y', 'o', 'e' y 'ni', que aparecen entre los primeros 32 segmentos ('u' ocurre en el núm. 69), están concentradas al principio de la tabla aquellas preposiciones que tienden a unir elementos al interior de sintagmas nominales ('de', *leche de cabra*; 'con', *pan con mantequilla*) o de perífrasis verbales ('a', *voy a comer*; 'para' y 'pa',

ver para creer), así como sus contracciones con artículos ('del', 'al'). También hay nexos oracionales ('que', 'como' y 'cuando'). Además, desde muy cerca del principio de la tabla están las formas verbales más importantes del verbo copulativo 'ser' (núm. 10 'es', núm. 28 'son', núm. 47 'fue', núm. 51 'era', núm. 52 'fuera' y núm 57 'será'). Al igual que en las otras tablas, mientras más decrece el índice de ordenamiento, menos es obvio el carácter de nexo de los segmentos. Sin embargo, hay que notar que varios de ellos (si bien no los de mayor índice) son, más que nexos, partículas invariables con cierta independencia en sus contextos; es decir, son segmentos que pueden funcionar como sintagmas nominales ('todo', 'esto', 'aquí', 'allí', 'ahí', 'uno', 'algo', 'ahora', 'usté', 'usted') o adverbiales ('también', 'todavía', 11 adverbios en *~mente*, etc.) en diferentes posiciones de la oración, prácticamente en cualquiera, cosa que explica la similitud de sus cliticidades.

Lo importante es que en todas estas tablas los valores mayores de los índices de ordenamiento corresponden a las formas que le dan estructura al discurso, ya sea matizando el contenido de palabras plenas, conectándolas o asignándoles alguna función gramatical específica. Desde esta perspectiva cuantitativa, sin embargo, no es muy clara la frontera entre los signos gramaticales y los de contenido. Y, dada la variación social, geográfica y estilística del lenguaje, por no hablar de su naturaleza cambiante, no tiene por qué ser de otra manera. En este contexto, los clíticos son sólo uno de los tipos de este conjunto difuso y constituyen en sí un grupo sin fronteras exactas.

3.7 Los clíticos en el *CEMC*

En esta sección se analizan los vocablos gráficos que el procedimiento identificó como los más aptos de ser clíticos, aunque en muchos casos tengan valores más bien bajos. La idea es examinar brevemente algunas de sus características, tales como su comportamiento morfológico y sintáctico, así como algunas particularidades fónicas pertinentes.

Como se mencionó arriba, no todo lo que ocurre en las tablas 3.5 y 3.6 (de la sección anterior) merece calificarse de clítico (especialmente en cuanto a los enclíticos se refiere). Allí se hizo evidente que ser clítico es una condición relativa, ya que la frontera entre los clíticos y lo demás es borrosa. Por eso, vale la pena examinar la naturaleza de lo que se reunió allí.

Con respecto a las características fonológicas de los segmentos en esas tablas, el que se trate o no de partículas átonas es una cuestión de conocerlas previamente. Muy aparte de que, por cuestiones pragmáticas, en ciertos enunciados estas partículas puedan percibirse, como se dijo arriba, con cierta intensidad accesoria (es decir, como tónicas), lo que nos incumbe es determinar cuáles son regularmente átonas. El problema es que no hay marcas de atonicidad en la escritura del español. De hecho, aunque podemos jactarnos de tener un sistema de escritura con acentos gráficos, la atonicidad no se puede inferir sólo a partir de la palabra gráfica. De todas maneras, es alentador constatar que lo que conocemos como partículas átonas se concentra en las primeras posiciones de ambas tablas.

Otra particularidad fónica pertinente, evidente en la palabra escrita, es la longitud en letras de los segmentos gráficos, que con ciertas transformaciones (como algunas de la tabla A.7 del apéndice), corresponde a la longitud en fonemas. Si se espera que los clíticos estén consti-

tuidos por una menor cantidad de fonemas o letras, podemos corroborar que en nuestra muestra las secuencias con menor longitud tienen menores rangos. Sin embargo, la preposición 'de' y la conjunción 'o', que son de las más cortas, tienen rangos relativamente mayores (respectivamente núm. 82 de la tabla 3.5 y núm. 80 de la 3.6). Esto significa que, aunque cliticidad y poco material fónico vayan de la mano, hay formas considerablemente más largas con más cliticidad, es decir, con menores rangos que estas dos. Así, aunque el desgaste fonológico parece resultado lógico de convertirse en clítico, no sólo hay desgaste cuando hay cliticidad. De todos modos, se observa en nuestra muestra la tendencia de los segmentos más reducidos a encabezar las tablas.

En cuanto a la estructura morfológica de los segmentos, empecemos con los "enclíticos". En la tabla 3.9 se repiten los primeros treinta. Aunque, como ya se dijo, no encontremos allí los enclíticos clásicos del español, se ve que por lo menos los primeros quince no son analizables en morfemas[24]. Además, en cuanto a su comportamiento morfológico, si omitimos adverbios y adjetivos, nos quedamos sobre todo con preposiciones, que son los morfemas imprescindibles de —valga la redundancia— el sintagma preposicional.

Pero precisamente la aparición de las preposiciones en esta tabla bien podría resultar intrigante, sobre todo si estamos acostumbrados a representar y organizar la estructura interna de sintagmas mediante esquemas arbóreos independientes de contexto como el de la figura 3.6.

En este esquema se especifican relaciones tanto entre la preposición y la frase nominal

[24]La estructura morfológica del fósil 'conmigo' no es transparente para el hablante promedio del español, es decir, en nuestra sincronía se trata de un solo morfema.

Tabla 3.9: Las 30 formas más "enclíticas" del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
1.	en	50123	0.3852	0.8568	0.4716	0.404
2.	a	45525	0.4763	0.6961	0.2198	0.153
3.	con	18747	0.4609	0.6525	0.1916	0.125
4.	por	19835	0.4012	0.6028	0.2016	0.1215
5.	y	60303	0.8478	0.9515	0.1037	0.09864
6.	sobre	2740	0.2413	0.4283	0.1869	0.08005
7.	hacia	911	0.2215	0.3945	0.173	0.06823
8.	entre	2528	0.2846	0.42	0.1353	0.05684
9.	durante	1017	0.2142	0.3682	0.154	0.0567
10.	ante	621	0.2038	0.3515	0.1478	0.05194
11.	conmigo	251	0.2067	0.3267	0.1201	0.03923
12.	contra	675	0.229	0.3365	0.1075	0.03619
13.	nada	2730	0.2346	0.3401	0.1055	0.03586
14.	antes	1288	0.2214	0.3224	0.101	0.03257
15.	desde	1904	0.253	0.3389	0.08595	0.02913
16.	social	503	0.226	0.3138	0.08782	0.02756
17.	nacional	445	0.2513	0.3319	0.08066	0.02678
18.	bajo	627	0.245	0.3235	0.07855	0.02541
19.	hasta	3018	0.314	0.3791	0.06515	0.0247
20.	natural	261	0.2049	0.2889	0.08397	0.02425
21.	libre	280	0.2297	0.3082	0.07843	0.02417
22.	mexicana	265	0.2306	0.3082	0.07762	0.02393
23.	mucho	1754	0.3214	0.3823	0.06092	0.02329
24.	total	483	0.2082	0.2882	0.08001	0.02305
25.	internacional	159	0.2069	0.2863	0.07937	0.02272
26.	después	2051	0.2224	0.2925	0.07014	0.02052
27.	juntos	103	0.2016	0.2752	0.07367	0.02027
28.	del	18621	0.6578	0.6869	0.02906	0.01996
29.	real	181	0.2098	0.2785	0.06867	0.01912
30.	para	14655	0.5623	0.5935	0.03119	0.01851

como entre dicha preposición y el nodo del que *cuelgue* el sintagma preposicional. A partir de esta aproximación se podría presumir que la relación con el sintagma nominal sería más fuerte. porque la preposición constituiría una capa de material que envuelve al sustantivo. Visto así, sorprende que las preposiciones aparezcan entre los "enclíticos". Sin embargo, si bien no es del todo extraño que se inserten algunas estructuras parentéticas entre la preposición y el sintagma nominal, su relación con lo que le antecede es más variada (sintagmas de todo tipo). Seguramente por eso las diferencias entre cliticidades de todas las preposiciones en la tabla 3.9 son menores a las cliticidades que representan sus asociaciones a la derecha (hacia los

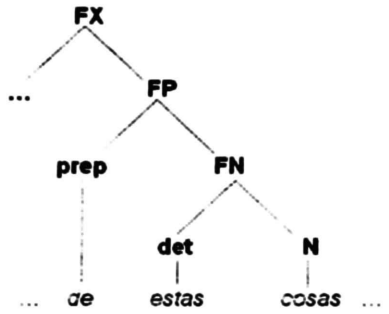

Figura 3.6: Esquema arbóreo de frase preposicional

sintagmas nominales que anteceden). En otras palabras, el que aparezcan aquí como enclíticos oculta sus relaciones hacia lo que les sigue y destaca sus enormes asociaciones con lo que les antecede. Así, aunque las preposiciones tienen una fuerte asociación con la clase abierta de segmentos de contenido que preceden —sustantivos—, no dejan de *colgarse* de algo, que generalmente les precede, y que puede ser casi cualquier clase abierta —sustantivos ('**Pedro** de Icaza', '**casa** de campo'), verbos ('**partir** a Madrid', '**andar** por la calle', '**poner** en la mesa') e, incluso, adjetivos ('**rojo** de envidia'). Esto asegura relaciones altamente entrópicas y económicas también hacia lo que les antecede.

El esquema de la figura 3.6 no refleja el *peso* de la relación del sintagma preposicional con su contexto sintáctico ni del sintagma con su preposición[25]. De todos modos, conviene tomarlo en cuenta porque esta situación representa un punto intermedio entre morfología y sintaxis: si por un lado se trata de la asociación morfológica de la preposición con su sintagma, por el otro, se deja ver la relación sintáctica del sintagma con estructuras superiores. Y, aunque según los datos hay una asociación ligeramente mayor con éstas últimas (finalmente, como

[25]Esto no tendría porque ser así. Ese peso podría ser la cuantificación de relaciones a partir de un corpus. Además, no sería tan descabellado asociarle a cada relación gráfica donde haya palabras función un peso típico de adhesión cuantitativa. Esto sería especialmente interesante en la representación de relaciones morfológicas al interior del sintagma e, incluso, de la palabra núcleo del sintagma.

se aprecia en la tabla 3.9, las diferencias entre cliticidades de cada lado de cada preposición —con excepción de 'en'— no es tan grande), no deja de destacar el hecho de que sólo dos pasen como proclíticos, los que casualmente suelen ser, con otros significados, también prefijos ('sin' y 'de', núms. 46 y 82 de la tabla 3.5). Sería interesante averiguar si en otros córpora se observa una tendencia similar.

Otra cuestión de carácter sintáctico se refiere al comportamiento de las categorías representadas en la tabla 3.9 (y más claramente en la versión larga de la tabla 3.6, página 210). A primera vista notamos que la mayoría de los 100 segmentos gráficos representa vocablos cuyo comportamiento sintáctico contrasta con el que esperaríamos de un clítico, especialmente porque pueden fungir como sintagmas plenos. Por ejemplo, a partir del núm. 16 (y después de haber descartado las preposiciones) ocurre una variedad de adjetivos y, más adelante, adverbios con una vida sintáctica propia. También están algunos pronombres: 'todos' (núm. 40), 'algo' (núm. 60), 'uno' (núm. 68) y 'toda' (núm. 70) de comportamiento similarmente libre. Ya hacia el final está la conjunción 'o' y los verbos finitos 'llegó', 'iba' y 'llega' (núms. 84, 90 y 91) con valores de encliticidad tan bajos que apenas vale la pena mencionarlos.

Como bien se sabe, los adverbios pueden ocurrir prácticamente en cualquier lugar de la oración, al principio o al final, entre sujeto y predicado, al interior del sintagma nominal o verbal. Además, si tomamos en cuenta su magnitud de fuerza adhesiva (véase la columna de la diferencia de cliticidades) observaremos que está en el orden de 0.06 (ver el vocablo 'mucho', único adverbio —aunque también adjetivo— de la tabla 3.9), dato que permite descartarlos como enclíticos verdaderos.

Cabe notar, sin embargo, que si bien se trata en su mayoría de vocablos adjetivos y

adverbiales, es decir, que pertenecen a dos de las categorías sintácticas tradicionales. el comportamiento sintáctico en español de por lo menos los adjetivos no es en realidad tan flexible como el de otras categorías. No en balde el lugar típico del adjetivo en esta lengua es después del sustantivo. Además, no cualquier cosa se puede insertar entre sustantivos y adjetivos, de hecho, sólo lo que puede ocurrir al interior de un sintagma adjetivo (**'el remedio común'** → **'el remedio *muy* común'** → **'el remedio *natural y muy* común'**). La inserción de cualquier otra cosa rompe la continuidad del sintagma (***'el remedio *tomó* común'**). Por todo esto, hasta habría sido extraño que no ocurrieran tantos adjetivos en esta tabla.

También es interesante el caso de los pronombres, ya que los enclíticos típicos del español también lo son. De hecho, en ciertos contextos pueden intercambiarse por razones estilísticas o, como estrategia pragmática para, por ejemplo, enfatizar algo. Considérense los siguientes tres grupos de sintagmas:

1. comprarla, comprándola, cómprala, cómprela, cómprenla;
2. comprarla toda, comprándola toda, cómprala toda, cómprela toda, cómprenla toda;
3. comprar toda, comprando toda, compra toda, compre toda, compren toda.

Los tres grupos se analizan de distinta manera: el primero es una lista de sintagmas verbales, mientras que el segundo y el tercero son las mismas series acompañadas del vocablo 'toda', que en el último grupo es claramente un sintagma nominal. En el segundo, 'toda' puede analizarse, ciertamente, también como un sintagma nominal (donde el clítico del sintagma verbal funciona como catáfora), pero ése no es el único análisis posible. De hecho, se puede plantear que los dos primeros grupos son más similares entre sí. La diferencia entre el

primero y el segundo residiría en que mediante el segmento 'toda' se enfatiza la completitud de lo referenciado por el pronombre enclítico ~*la*, como si ambos constituyeran el argumento enclitizado del verbo.

Algo similar puede decirse de los otros pronombres que ocurren en la tabla de "enclíticos". De hecho, 'uno' y 'algo' no acompañan al enclítico ~lo, sino que alternan con él (la ocurrencia de algo así en un corpus implicaría necesariamente que ambos pronombres no comparten referente):

1. *comprarlo$_i$ uno$_i$, *comprándolo$_i$ uno$_i$, ...
2. comprar uno, comprando uno, ...

1. *comprarlo$_i$ algo$_i$, *comprándolo$_i$ algo$_i$, ...
2. comprar algo, comprando algo, ...

En cuanto al hecho de que 'toda' no alterne con el clítico, sino que lo acompañe, seguramente está relacionado con la condición especial de la familia de pronombres representada por 'todo' que, además de pronombres, fungen como adjetivos determinativos y que, singularmente, co-ocurren con otros determinativos (**'todas estas cosas'**). También es de notarse que entre el indefinido y el adjetivo determinativo del sintagma nominal no ocurre nada, como tampoco entre el verboide con enclítico y la partícula para enfatizar la completitud de lo referido por el pronombre. Otra cuestión interesante es que se puede agregar un adjetivo al final de esta estructura y mantener la sensación de que algo se está enclitizando: **'cómpralas todas frescas'**.

Pero el caso de 'todo' es peculiar porque para otros pronombres la alternancia con el clítico es, como con 'uno' y 'algo', obligatoria (**'comprando algunos'** *versus* ***'comprándolos$_i$ algunos$_i$'**). Aunque en nuestra muestra no estén todos los que podrían alternar con los enclíticos, es significativo que estén representados por 'toda', 'uno' y 'algo'. Lo importante es que, si la alternancia entre un enclítico y un sintagma pleno es posible, no parece tan casual la aparición de lo que actúa como ese sintagma en una lista de segmentos con cierta encliticidad.

Por todo esto, es revelador que estas formas (y no cualquier otra) sean precisamente las que ocurren en la lista de la tabla 3.6, con todo y sus bajos niveles de encliticidad. De hecho, no parece descabellado decir que si en un futuro algún nuevo enclítico ha de emerger en el español de México, la probabilidad de que provenga de entre los segmentos de dicha tabla no es despreciable.

Con respecto a los proclíticos, los treinta primeros de la tabla 3.5 se reproducen en la tabla 3.10. Examinemos primero su morfología. Nótese que al considerar las 100 formas se aprecia que la mayoría de las no analizables ocurren dentro de las primeras treinta. Si bien algunos de los de menor rango también son analizables, como los artículos o pronombres con marca de género o número (por ej. l-a-s, l-o, le-s), los no analizables abarcan por lo general los rangos menores ('se', 'su', 'un', 'el', 'me', etc.; núms. 1, 4, 5, 7, 9, 11). Es curioso que los últimos inanalizables de la tabla 3.5, 'luego' y 'pues' (núms. 98 y 97), se alejen considerablemente de lo que sería un proclítico típico.

Además, los más proclíticos son menos aptos de ocurrir como sintagmas plenos. Más bien, tienden a acompañar, con la función de matizar o complementar los contenidos discursivos,

Tabla 3.10: Las 30 formas más proclíticas del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
1.	se	33623	0.3807	0.9862	0.6055	0.5972
2.	la	73110	0.3575	0.9243	0.5668	0.5239
3.	los	31231	0.3121	0.8691	0.557	0.4841
4.	su	12520	0.3099	0.8538	0.5439	0.4643
5.	un	19765	0.3766	0.8745	0.4979	0.4354
6.	una	16473	0.3718	0.8362	0.4645	0.3884
7.	el	51708	0.4439	0.8595	0.4156	0.3572
8.	las	20882	0.3144	0.732	0.4176	0.3057
9.	me	10410	0.2709	0.6888	0.4179	0.2878
10.	sus	5267	0.2826	0.6948	0.4122	0.2864
11.	le	8502	0.2552	0.617	0.3618	0.2232
12.	nos	3399	0.2711	0.5977	0.3266	0.1952
13.	te	3389	0.2425	0.5413	0.2988	0.1617
14.	ser	2830	0.2356	0.5262	0.2906	0.1529
15.	les	2021	0.2368	0.5156	0.2788	0.1438
16.	esta	2925	0.2673	0.5261	0.2588	0.1361
17.	lo	13705	0.3125	0.5455	0.233	0.1271
18.	más	9778	0.4075	0.6127	0.2052	0.1258
19.	dos	3216	0.282	0.5186	0.2366	0.1227
20.	muy	4915	0.3673	0.5758	0.2085	0.1201
21.	esa	1666	0.2703	0.4936	0.2233	0.1102
22.	había	2023	0.2537	0.4733	0.2196	0.1039
23.	al	11179	0.5579	0.6966	0.1387	0.09664
24.	ese	1977	0.2666	0.4693	0.2027	0.09511
25.	está	4108	0.2983	0.4914	0.1931	0.09486
26.	puede	2480	0.2679	0.4671	0.1992	0.09303
27.	otros	1499	0.2211	0.434	0.2129	0.09238
28.	tan	1591	0.3544	0.5245	0.1701	0.08919
29.	mi	4840	0.284	0.4695	0.1855	0.08709
30.	están	1561	0.2985	0.477	0.1785	0.08515

a otros que sí podrían actuar como sintagmas completos[26]. Lo interesante es que la mayoría de los segmentos que ocurren en la tabla 3.10 típicamente preceden sintagmas de algún tipo y todo lo que ocurre entre los núcleos de esos sintagmas y cada segmento es material del mismo sintagma. Por eso, muchos se pueden considerar parte de los sintagmas que preceden:

1. tan <adj.> → <adj.>

2. esta <s.> → <s.>

[26]Si bien 'luego' y 'pues' pueden ocurrir solos (por ej., en respuestas cortas), fuera de contexto no dicen nada. Además, según la evidencia cuantitativa, parecen anteceder cosas mucho más que cerrar enunciados.

3. no se lo <v.> → no se <v.> → no <v.> → <v.>

Nótese que el adverbio 'no' cumple con estas características. Si no aparece entre los primeros treinta (de hecho, tiene rango 40), es porque los datos cuantitativos representan a dos segmentos idénticos en forma, pero diferentes en distribución: el 'no' que ocurre solo (por ej. como respuesta), y que no se considera clítico, y el proclítico del sintagma verbal. Por ejemplo, en **'no, no le dice nada'**, el primer 'no' constituye un sintagma adverbial, mientras que el segundo es clítico del verbal.

Con respecto a su función dentro de los sintagmas, es alentador que la primera parte de la tabla esté densamente poblada, como se mencionó en la sección anterior, de los proclíticos tradicionales del español, así como de adjetivos determinativos de varios tipos y de ciertos adverbios muy frecuentes pero de comportamiento restringido al interior de los sintagmas ('muy', 'tan', etc.). En general, las formas de comportamiento sintáctico restringido ocurren al inicio (tablas 3.5 y 3.10). Así, los pronombres proclíticos siempre preceden al núcleo verbal: los determinativos definidos e indefinidos, así como los posesivos ('su', 'un', 'sus', 'esta', 'mi', 'tu', 'mis', etc., núms. 4, 5, 10, 16, 29, 33, 56) anteceden al núcleo sustantivo. Los adverbios son los que típicamente preceden adjetivos: 'muy', 'tan', 'más', 'menos' (núms. 20, 28, 18, 61). Mientras menor es el rango de los segmentos, menor libertad sintáctica y mayor es su calidad de proclíticos; mientras mayor es el rango, más libres sintácticamente y menos justificable su pertenencia a una lista de clíticos.

Sin embargo, lo interesante es, como en el caso de los enclíticos, lo que menos se ajusta a la noción tradicional de clítico; por ejemplo, los formas verbales finitas. En la sección anterior se apuntó que ciertos verbos se adhieren a otros de contenido pleno para formar nuevos paradig-

mas verbales. De hecho, normalmente no ocurren solos y sólo por cuestiones estilísticas o pragmáticas cambian su lugar habitual con respecto al verbo de contenido con el que se combinan (nótese la afectación en los siguientes ejemplos, debida a la permutación de los verbos: 'la sopa comido ↔ **había**', 'nunca hablar ↔ **puede**', 'ayer llover ↔ **hizo**' y 'aquí cantar ↔ **quería**') o en relación con los sustantivos con que forman unidades fraseológicas ('hambre ↔ **tienen**', 'las gracias ↔ **dar**', etc.)[27]. En este sentido, están restringidos sintácticamente.

Pero algunas formas finitas de otros verbos, como 'quedar' y 'estar', que esperaríamos ocurrieran en cualquier posición de la oración también aparecen como proclíticas. El verbo 'quedar', cuyas formas ocurren hacia el final de la tabla (núms. 72 y 90), es interesante porque es intransitivo o pronominal y se esperaría que ocurriera en cualquier lugar. Pero al examinar el corpus se ve que típicamente va seguido de sus sujetos (**'quedó *la cruz de milpa*'**) o de algún complemento circunstancial (**'Charo quedó *como difunta*'**). Por otra parte, el verbo 'estar' todavía es más interesante porque sus formas ocurren desde el principio de la tabla (esto es, aparecen como muy proclíticas): las finitas de 3ªp. presente y copretérito ('está', 'están', 'estaba', 'estaban' y el subjuntivo 'esté'; núms. 25, 30, 57, 80 y 78) y su infinitivo (núm. 58). Al considerar que en perífrasis verbales preceden gerundios (**'estaba *comiendo*'**), forman predicados adjetivos (**'está *bonita*'**) o, incluso, introducen sintagmas nominales (**'está *una señora*'**), donde se expresa la presencia del sujeto pospuesto, por lo que funge como existencial ($\exists x|x$ es señora). Su uso como intransitivo con sujeto antecedido o implícito ('[SN] está') no parece ser el más representativo. Dado esto, sería sorprendente que no aparecieran en una lista de formas gráficas con fuerte asociación hacia lo que ocurra

[27] Esto no excluye que ciertos verbos, cuyas formas finitas tienden a aparecer hacia el final de la tabla 3.5, [illegible] en ciertas construcciones como verbos únicamente de contenido: 'quería dinero', 'tienen un [illegible]'

a la derecha.

Como hemos visto, más allá de algún hipotético umbral entre lo clítico y lo no clítico, entre los resultados de este capítulo no ocurre simplemente cualquier vocablo gráfico del corpus. Más bien, aparecen formas específicas que exhiben por lo menos una de las características, muy a menudo incipientemente, de los segmentos más cliticales. En otras palabras, al haber ciertos patrones comunes entre los segmentos menos clíticos, los que quedan fuera de esa frontera difusa, no se puede confirmar la intuición inicial de que lo que no es clítico es ruido, es decir, que sea un conjunto heterogéneo de formas fuera de cualquier orden.

Hasta aquí sólo se han analizado los segmentos más aptos de ser clíticos. Es obvio que falta mucho que considerar. Al final del próximo capítulo examinaremos una versión más completa de los más gramaticales (tabla 3.7). En concreto, se analizarán los 500 vocablos consignados en la tabla D.1 del apéndice (página 423).

3.8 Observaciones finales

En este último apartado se hace un resumen del capítulo. En esencia se exploró la manera de aplicar el método de descubrimiento de afijos del capítulo anterior en el nivel de la palabra gráfica. No es de sorprenderse que los criterios para descubrir afijos también se puedan aplicar al descubrimiento de clíticos. De hecho, si hemos de concebir, siguiendo a Meillet, afijos y clíticos como dos grados diferentes de fusión de unidades originalmente independientes, podemos suponer que todas aquellas palabras con más función gramatical que los vocablos de contenido se pueden descubrir mediante métodos similares. Así, en este experimento se

estimaron las cantidades de entropía y economía inherentes a cada vocablo gráfico del *CEMC*.

Si bien por un lado se omitieron tanto la frecuencia como el número de cuadros[28] en el cálculo de cliticidad, por el otro se exploró otro índice basado en la puntuación. Esta medida adicional —aunque también puede concebirse como resultado y no característica del carácter de clítico que puedan tener los segmentos de un corpus— funciona efectivamente en la determinación de fronteras sintagmáticas y es muy fácil de calcular: la probabilidad de ocurrencia de signos de puntuación entre segmentos. Esto implica la concepción de por lo menos dos tipos de cliticidad: una general que no toma en cuenta la puntuación y otra "escrita" que utiliza los indicios de puntuación como suerte de resistencia opuesta a la cliticidad general. No se puede ignorar que esta llamada cliticidad general está de todas maneras impregnada de lengua escrita (la transcripción es inevitable) principalmente porque las fronteras gráficas de las palabras españolas no necesariamente coinciden con sus fronteras verdaderas. En el siguiente capítulo encontraremos esta misma situación con respecto a lo que llamaremos "glutinosidad" general y escrita y veremos que los tintes de lengua escrita también son inevitables en la estimación de una fuerza de adhesión entre segmentos, aun cuando no hay puntuación de por medio.

En cuanto al trabajo de este capítulo, los resultados del experimento nos dan los elementos para rechazar la hipótesis de cliticidad tal y como se presentó en la introducción y al principio de este capítulo. La cliticidad no es meramente una cantidad directamente proporcional a la información que fluye entre un clítico y un no clítico, ni a la economía presente al combinar un signo gramatical con uno de contenido. La cliticidad debe ser una función de la diferencia de

[28]Se omitieron principalmente porque parecen redundantes y no muy estables (por ej. exhibieron una fluctuación importante entre sufijos y prefijos del capítulo anterior).

información que pueda fluir de un lado o de otro hacia lo que ocurra alrededor del supuesto clítico y, simultáneamente, de la desigualdad entre las relaciones económicas que el signo establece con los signos que lo preceden y las que entabla con los que le anteceden. Si estas dimensiones no se restan, no tenemos la cantidad de cliticidad del signo, sino una medida de su inserción en ese sistema que es la lengua.

Capítulo 4

Glutinometría en el *CEMC*

> Si el lenguaje es un edificio y si los elementos significantes del lenguaje son los ladrillos de que está hecho ese edificio, entonces los sonidos del habla no pueden compararse sino con el barro, todavía sin moldear y sin cocer, con el cual se fabrican los ladrillos. **Edward Sapir**

> What can be observed either directly or indirectly is a set of traits of some concrete system. If observation is to be precise it must be quantitative because concrete systems have quantitative properties, if only because they exist in determinate amounts and in spacetime. Quantitative observation is measurement. Whenever numbers are assigned to certain traits on the basis of observation, measurements are being taken. **Mario Bunge**

Dado que, como vimos en los capítulos anteriores, tanto la afijalidad como la cliticidad pueden cuantificarse mediante combinaciones de diversas dimensiones medibles ya sea entre segmentos de palabras o entre palabras, en este capítulo se examinan sus semejanzas y diferencias para intentar una generalización de estos conceptos[1]. En la primera parte, después de los antecedentes (donde se busca mostrar que las ideas presentadas aquí no son del todo nuevas en la lingüística), se investiga primero la relación entre la afijalidad y la cliticidad.

[1]La primera parte de este capítulo está basada en la ponencia "Propiedades lingüístico-cuantitativas de cadenas de caracteres (segmentos, palabras, vocablos) en córpora de lenguajes naturales: *afijalidad y cliticidad* en el español de México" [98], VI Congreso Internacional de Lingüística en el Noroeste, Hermosillo, Sonora, diciembre, 2000.

Luego se examinan los problemas que surgen al aplicar estos conceptos en la determinación de fronteras entre sintagmas y, por último, las posibles manifestaciones de todo esto en el tiempo. Dicho de otra manera, se analiza la mecánica de las relaciones entre estas medidas para definir una hipotética fuerza de enlace, pegajosidad formal o *glutinosidad*[2] entre los elementos del discurso. En seguida, se presentan los conceptos básicos de la teoría de la medición, como marco para hacer operativa esta noción de *glutinosidad* y construir un esquema *glutinométrico* formal y coherente. Por último, se presentan los resultados de la aplicación de estas reflexiones en el *Corpus del Español Mexicano Contemporáneo.*

4.1 Antecedentes

En esta breve sección se revisan algunas intuiciones que apuntan a la existencia de una fuerza de enlace formal o *glutinosidad* como fenómeno lingüístico general. Se trata de ideas muy generales que se han abordado indirecta, casual o, incluso, metafóricamente, pero que es pertinente tener presente aquí para establecer la necesidad de elaborar dentro de la lingüística este tipo de construcciones teóricas.

La presunción de la existencia de una fuerza de atracción entre las unidades léxicas de una lengua no se limita a una sola lengua, sino que pretende rendir cuenta de un fenómeno común a todas las lenguas. La búsqueda de los universales del lenguaje no es nada nuevo ni exclusivo de alguna corriente de pensamiento en particular. De hecho, especialmente en el siglo XVII, antes del nacimiento oficial de lo que hoy conocemos como lingüística, floreció

[2]Del adjetivo *glutinoso* (lat. *glūten*, 'cola, engrudo'; véase Corominas y Pascual, *op. cit.* [44] 1991, *s.v.* GLUTEN), aquí se referirá a una propiedad no tanto *de* los elementos (como se ha concebido, por comodidad, la afijalidad y la cliticidad en los capítulos anteriores), sino más bien *entre* elementos.

la distinción entre la gramática general[3], como *ciencia*, y las gramáticas particulares, como artes gramaticales (*grammaire générale vs. grammaires particulières*).

Sobra decir que los procedimientos para determinar lo universal de las lenguas varían en gran medida, pero en esencia se pueden localizar entre dos polos metodológicos de carácter filosófico: del racionalismo introspectivo al empirismo extremo. En el primero se descuidan los datos reales y en el opuesto se presume que todos los fenómenos lingüísticos en todos sus aspectos son directamente observables. Entre los dos polos podemos concebir toda una gama de métodos posibles en el estudio del lenguaje (ya sea que se trate de determinar universales o particulares) en los que se apliquen alternativamente procedimientos más cercanos a un polo u otro.

Para determinar un método descriptivo de todas las lenguas es indudable que no podemos prescindir de ninguno de los dos, pero es necesario ser cuidadoso en el orden de aplicación de los procedimientos de introspección y de observación o experimentación: no conviene depender más de la introspección para conocer a los fenómenos del lenguaje, pero tampoco podemos quedarnos sin generalizar o razonar la evidencia de lo particular en la determinación de los universales.

Tal vez la alusión más antigua que nos quede de una reflexión general sobre el carácter económico del lenguaje sea la de Varrón, quien, algunas décadas antes de Cristo, se percató de la carga que le significa a la memoria no tener afijos derivativos y flexivos, cosa que expresó clara y elocuentemente:

---

[3]Entre los pensadores que Chomsky luego designó bajo el término de 'gramáticos cartesianos': véase *op. cit.* [35] 1966. pp. 52-54.

> La "declinación" [tanto de nombres y adjetivos como de conjugación verbal] se ha aplicado no sólo a la lengua latina, sino a la de todos los hombres, por una razón útil y necesaria. De no haberlo hecho así no podríamos aprender un número tan grande de palabras (ya que las formas naturales en que los vocablos se declinan son infinitas), y aunque las hubiéramos aprendido no podríamos descubrir a partir de ellas qué sistema [sic] las relaciona entre sí[4]. En cambio, ahora sí podemos percibirlo porque se trata de algo semejante, de algo que ha derivado: [..] Dos son, en general, los orígenes de las palabras: la imposición y la flexión. La primera viene a ser la fuente; la segunda, el río. Los hombres quisieron que las formas "impuestas" fueran las menos posibles, con el fin de aprenderlas cuanto antes; y que las "flexionadas" fueran el mayor número posible, para que todos pudiesen emplear aquellas que fuera necesario utilizar[5] .

Muchos siglos después, entre los lingüistas mecanicistas estadounidenses, para quienes era natural presumir que los fenómenos, sus causas y sus efectos son observables, el trabajo con córpora fue primordial. En especial, esquemas como el cálculo de entropías para determinar el contenido de información son muy atractivos en este marco: hay una fuente de información, un transmisor que la codifica y emite señales, luego hay señales recibidas y decodificadas por un destinatario que selecciona un mensaje entre varios posibles[6]. De hecho, la teoría de la comunicación de Shannon y Weaver se presenta en un marco de tres niveles de problemas: el técnico, el semántico y el de eficacia (*effectiveness*), el último de los cuales se refiere a la efectividad con que se afecta una conducta mediante un mensaje. La contribución de su trabajo está, por supuesto, en el nivel técnico, pero los alcances de su marco general no se han explorado suficientemente sobre todo en lo que al lenguaje natural se refiere.

---

[4]En la versión inglesa de Kent: "for if this system [sic] had not developed, we could not learn such a great number of words as we should have —for the possible forms into which they are inflected are numerically unlimited— nor from those which we should have learned would it be clear what relationship existed between them so far as their meanings were concerned" (Roland G. Kent. *op. cit.* 1938. p. 373); del original latino: *nisi enim ita esset factum, neque discere tantum numerum verborum possemus (infinitae enim sunt naturae in quas ea declinantur) neque quae didicissemus, ex his, quae inter se rerum cognatio esset, appareret.*

[5]Varrón, *De lingua Latina* [133], tr. Manuel-Antonio Marcos Casquero, VIII, 3-5, pp. 292-295.

[6]Weaver, art. cit. [134] 1964, p. 7.

Podría presumirse que el cálculo de índices de economía y entropía en el estudio de córpora sería una tarea sobre todo mecanicista, que presume que el fenómeno lingüístico ocurre completo en un corpus y por lo tanto es posible observarlo allí en su totalidad. Es decir, que todo lo necesario para medir qué tan económica o entrópica es una estructura está presente allí. Lo mismo se puede decir de los cuadros y la puntuación: todo está allí para que lo contemos, para que sea la causa de lo que concebimos como una propiedad de las estructuras observadas.

Sin embargo, ninguna glutinometría se podría dar el lujo de postularse como retrato completo de los fenómenos lingüísticos; es decir, no puede asumir que no falte en el corpus nada pertinente a los fenómenos que mida: desde factores lingüísticos hasta extralingüísticos. pasando por los procesos mentales más diversos, que sin manifestarse abiertamente en un corpus, podrían muy bien dejar huellas de las más variadas índoles e, incluso, causar ruido considerable en el cálculo de cualquier glutinosidad hipotética. Pero presumir que el corpus proporciona información suficiente para conocer la lengua allí representada no era un gran problema para los descriptivistas estadounidenses (que Chomsky criticara tanto precisamente por eso), por lo que seguramente un esquema de medición de fenómenos lingüísticos les habría parecido interesante. Nótese, por ejemplo, el entusiasmo de Hockett por la teoría de la información en su reseña del libro de Shannon y Weaver[7].

De hecho y como vimos en el primer capítulo, Harris —al contar los fonemas que le siguen a un segmento— está midiendo esta energía de adhesión. La variedad de fonemas que pueden ocurrir después de una cadena de caracteres es directamente proporcional a la fuerza

[7]Hockett [67]. *Language* 29(1953), pp. 69-93.



Figura 4.1: Fronteras entre morfemas en la oración *Dogs were indisputably quicker.*

Figura tomada de Harris (*A Theory of Language and Information* [66], Clarendon, Oxford. 1991, p. 172). Los límites entre segmentos fueron marcados por Harris mediante puntos. Lo interesante es notar que las cuentas altas de fonemas (v) corresponden en gran medida a estas marcas. La única excepción está en el punto entre 'ab' e 'ly' que alcanza apenas un número de 4 fonemas ($v = 4$).

de adhesión entre esa cadena de caracteres y los segmentos que le siguen. Como se ve en la figura 4.1, mientras más variedad de fonemas, más indicio de frontera entre segmentos. lo que implica una mayor fuerza de unión entre éstos.

Entre aquellos lingüistas para los que el lenguaje no se puede caracterizar exclusivamente en términos de causas y efectos visibles al investigador, también se pueden encontrar intuiciones que apunten a una energía glutinosa. Así, en *El lenguaje*, Sapir imagina una atracción entre palabras y elementos que se han expresado en cierto orden y que puede solidificarse en grupos de elementos y formar una sola masa, que aun después de cristalizada puede volverse a aflojar:

> Así, pues, las palabras y los elementos, una vez que se han expresado en cierto orden, no sólo tienden a desarrollar algún tipo de relación entre sí. sino que son atraídos más o menos el uno al otro. Se puede presumir que precisamente este "más o menos" es lo que, en resumidas cuentas, da origen a aquellos grupos de elementos, firmemente solidificados (elemento o elementos radicales más uno o

> más elementos gramaticales), que hemos estudiado como palabras complejas. Con toda verosimilitud, no son sino series de elementos que se han contraído, formando una sola masa, a partir de otras series, o de elementos aislados en la corriente del habla. Mientras están plenamente vivos, o, dicho en otras palabras, mientras son funcionales en cada punto de su estructura, pueden mantenerse a una distancia psicológica de sus vecinos. A medida que van perdiendo su vida individual, caen en brazos de la frase en cuanto conjunto, y la serie de las palabras independientes vuelve a adquirir la importancia que había transferido, en parte, a los grupos cristalizados de elementos. De esta manera, el lenguaje está apretando y aflojando sin cesar sus concatenaciones de palabras. En sus formas más sintéticas (como en latín o en esquimal), la "energía" de la secuencia queda encerrada, en gran parte, en complejas formaciones de palabras, viene a transformarse en una especie de energía potencial que quizá no se libere durante milenios. En sus formas más analíticas (como en chino o en inglés), esta energía es móvil, pronta para ser empleada en el servicio que se exija de ella.[8]

Es de notarse que Sapir mismo concibe esto —aunque sea metafóricamente— como una energía, una potencia que crece y decrece a través de los milenios. Esta es precisamente una de las acepciones, la más antigua, del concepto de 'gramaticalización'[9]. De hecho, Meillet se refirió a esta idea como la evolución de las formas gramaticales (afijos, palabras gramaticales, etc.) a partir de palabras de contenido o formas léxicas[10]. Mientras más se desgasta fonológicamente y semánticamente una forma, mayor es la fuerza con que las palabras accesorias se pegan a las palabras principales:

> L'affaiblissement du sens et l'affaiblissement de la forme des mots accessoires vont de pair; quand l'un et l'autre sont assez avancés, le mot accessoire peut finir par ne plus être qu'un élément privé de sens propre, joint à un mot principal pour en marquer le rôle grammatical.

---

[8]Sapir *op. cit.* [122] 1992 [1921], p. 131.

[9]Para Lázaro Carreter (*op. cit.* [91] 1990) es la única acepción: gramaticalización es un "proceso mediante el cual una palabra se vacía de contenido significativo, para convertirse en mero instrumento gramatical". Sin embargo, Bußmann (*op. cit.* [28] 1990, *s.v.* GRAMMATIKALISIERUNG.) consigna otras acepciones relacionadas con los aspectos semántico-pragmáticos del fenómeno (debidas a varios lingüistas, como Heine, Traugott, Bybee, Givón, etc.), que aquí no tomaremos en cuenta.

[10]Véase Meillet, art. cit. [101] 1912 en *op. cit.* [100] 1958, p. 139.

Como se verá a continuación, hay varias maneras de concebir y, por lo tanto, estimar esta energía entre segmentos de un corpus.

4.2 La lógica del esquema de *glutinosidad*: hacia un índice cuantitativo

En esta sección se examinan las semejanzas y diferencias en la mecánica de las relaciones entre las dimensiones involucradas tanto en el cálculo de la afijalidad de segmentos, como en el de la cliticidad de palabras (en el sentido general planteado en la introducción[11]) para cuantificar la fuerza de enlace entre estos elementos del discurso (entre segmentos de palabras y entre palabras).

Como se vio en los capítulos anteriores, la afijalidad de un segmento se puede concebir como una combinación de los índices de economía, cuadros y entropía medibles entre dos segmentos de palabra. La cliticidad, por otra parte, también se puede describir en razón de estas dimensiones, pero entre lo que hemos definido como palabras. Por un lado, para que un segmento sea afijal, se espera una segmentación económica, con un alto número de cuadros y una alta entropía (ya sea en una u otra dirección). Por el otro, el índice de cliticidad de las palabras se calculó esperando también una alta entropía y un índice alto de economía, además de tomarse en cuenta la aparición de caracteres no alfabéticos para calcular un índice de puntuación que debe ser mínimo entre un clítico y la palabra a la que

[11] De hecho, falta un término apropiado para designar esto, ya que, como vimos en el capítulo tres sobre el clítico, la cliticidad propiamente se entiende mejor como la diferencia entre las glutinosidades de cada lado de la palabra gráfica. A falta de un mejor término, en esta sección 'cliticidad' se referirá a una sola glutinosidad, la mayor, sin substracción de la otra.

se adhiere. Evidentemente, hay mucho en común en el cálculo de los índices de afijalidad y cliticidad. Sin embargo, resalta una diferencia aparente: el índice de afijalidad indica dónde segmentar una palabra, es decir, es la medida de *menor* asociación entre segmentos, mientras que la cliticidad indica qué palabras gráficas tienden a asociarse con otras (medida de *mayor* asociación). Esto podría llevarnos a hipotetizar que la afijalidad y la cliticidad sean fuerzas opuestas. Los dos subapartados siguientes se ocupan de analizar esta hipótesis. Luego, en el siguiente se examina el tipo de asociación que cabe esperar entre sintagmas y cómo las nociones involucradas en el cálculo de los índices de afijalidad y cliticidad podrían o no aplicarse en su estimación. Por último se presentan algunas reflexiones sobre el esquema propuesto con respecto a la diacronía del lenguaje.

4.2.1 La afijalidad y la cliticidad como fuerzas opuestas

Supongamos que la afijalidad y la cliticidad son dos manifestaciones opuestas de una misma idea. Es decir, si imaginamos, por un lado, un tipo de fuerza de enlace formal entre segmentos de palabras (formados por fonemas o grafemas) y entre palabras en una cadena hablada (o escrita) —una especie de pegamento estructural entre los elementos del discurso—, las fronteras afijales (y podríamos esperar que también las sintagmáticas) corresponderían a los puntos más débiles —donde el pegamento estructural es menor. Por el otro, los espacios de unión entre clíticos y palabras corresponderían a una pegajosidad relativamente mayor que aquella entre palabras plenas (y menor a aquella entre bases y afijos). Dicho de otra manera, al interior de un segmento que represente un afijo aislado o la raíz de una palabra, esta glutinosidad tendrá su valor mayor. Luego, entre morfemas al interior de la palabra

la fuerza sería menor; y entre clíticos y palabras sería mayor que entre palabras plenas y. finalmente, sería más débil entre sintagmas.



Figura 4.2: Glutinosidad más alta al interior de la palabra

Véase por ejemplo la figura 4.2. El área sombreada representa la glutinosidad. Dentro de cada palabra, esta pegajosidad aparece con sus valores más altos. Las hendiduras más profundas (donde hay menos glutinosidad) representan las fronteras entre sintagmas. Aquellas no tan profundas, el pegamento entre clíticos y palabras plenas y, las menos profundas. los puntos flacos dentro de las palabras correspondientes a las fronteras morfológicas. Como la cliticidad, esta glutinosidad puede medirse en función de, por lo menos, los conceptos formales que sustentan los índices de economía y entropía (y para una glutinosidad de la lengua escrita se puede tomar en cuenta la distribución de los signos de puntuación).

El experimento de descubrimiento de afijos del español, descrito en el capítulo primero, mostró que las fronteras entre fonemas al interior de raíces y de afijos (segmentaciones no mor-

fológicas) no exhiben niveles altos de afijalidad. Esa es precisamente la razón que nos permite presumir que allí no hay una división morfológica. Y si hemos de concebir la glutinosidad como más alta allí que en cualquier otro lado, entonces la podríamos medir mediante la diferencia entre el valor más alto posible (que en escalas normalizadas sería la unidad) menos la afijalidad calculada en esa segmentación (que en escalas normalizadas sería mayor que cero y menor que la unidad). Así, aunque al interior de una raíz esperemos simultáneamente índices mínimos de economía y entropía, podemos imaginarnos allí una glutinosidad máxima:

$$GL = 1 - AF$$

donde GL representa la glutinosidad entre los segmentos de un vocablo y AF es un índice normalizado (entre 0 y 1) que mide la afijalidad de uno de estos segmentos. Así, si un segmento es muy afijo, hay entre éste y la base poca glutinosidad, y, por el contrario, si su afijalidad no es alta, entonces hay allí más pegamento.

Por otra parte, ya que los datos nos muestran que los valores más altos de cliticidad corresponden a puntos de asociación alta entre vocablos, podemos concebir la glutinosidad como directamente proporcional a la cliticidad:

$$GL = CL$$

que indica que la cantidad de glutinosidad asociada a un clítico o palabra gráfica corresponde a la cantidad de cliticidad asociada a éste o ésta. Esto equivale a considerar a la afijalidad y la cliticidad como fuerzas opuestas. Sin embargo, como veremos a continuación, esta no es la mejor manera de concebirlas.

4.2.2 La afijalidad y la cliticidad como dos instancias de la misma fuerza

Sin embargo, hay que examinar con detenimiento las consecuencias de querer concebir la afijalidad y la cliticidad como dos fuerzas opuestas. El experimento del primer capítulo mostró, como se dijo arriba, que las fronteras entre fonemas al interior de raíces y afijos (segmentaciones no morfológicas) no exhiben niveles altos de afijalidad. Y el hecho de querer representar una glutinosidad máxima al interior de las raíces es la única razón para querer concebir la afijalidad como opuesta a la cliticidad. Pero esto no es necesario: el que no haya pegamento entre los fonemas al interior de un segmento no quiere decir que no estén unidos, sino que simplemente no necesitan pegamento para estar unidos. Esto significa, no que haya más pegamento donde hay menos afijalidad, ni más afijalidad donde menos pegamento, sino que en cada segmentación hay tanto pegamento como afijalidad. Aquí podemos aplicar la metáfora de Sapir del lenguaje como edificio, donde los elementos significantes son los ladrillos. No hay cemento entre los fonemas al interior de un morfema, porque forman parte del mismo ladrillo. Así, la cantidad de pegamento entre los morfemas de una palabra gráfica corresponde a la afijalidad, mientras que la cantidad de pegamento entre morfemas al exterior de la misma corresponde a la cliticidad, es decir, ambas son dos instancias de la misma fuerza glutinosa. En la figura 4.3 los valores de afijalidad y cliticidad se representan como fuerzas de la misma especie.

El área sombreada indica la cantidad de glutinosidad. Los picos más altos representan la cantidad de afijalidad que une los morfemas al interior de la palabra. Luego, los medianos corresponden a las uniones entre clíticos y palabras plenas. Finalmente, los picos menores

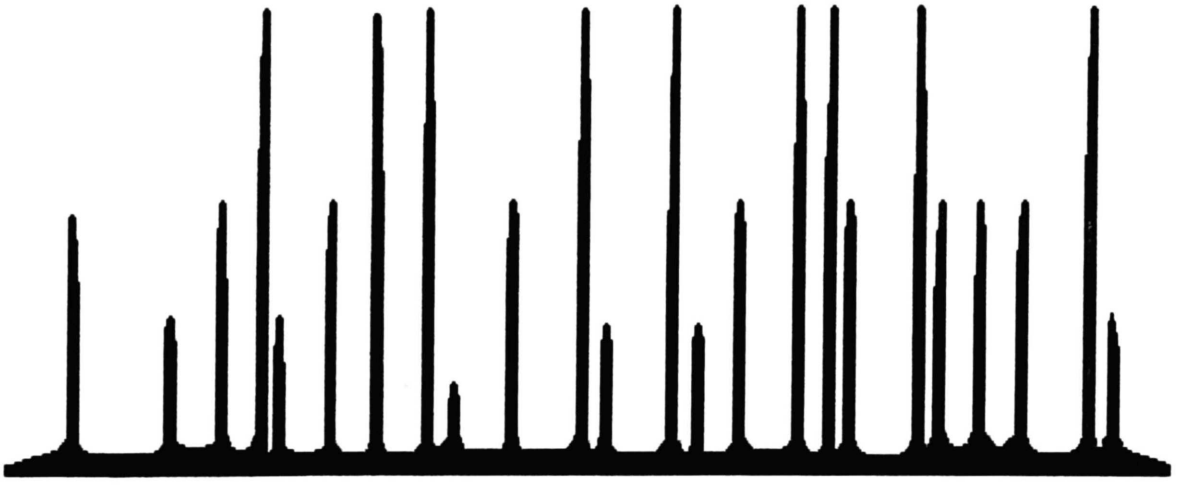

Figura 4.3: Afijalidad y cliticidad como instancias de la misma fuerza

representan las fronteras entre sintagmas.

De esta manera, a cada punto de una cadena de palabras, se puede calcular un índice de glutinosidad de la siguiente manera:

$$GL = \left\{ \begin{array}{lcl} AF & : & \text{al interior de la palabra} \\ CL & : & \text{al exterior de la palabra} \end{array} \right\} = kh$$

donde GL es la glutinosidad calculable a partir de índices de entropía (h) y economía (k), los cuales le son directamente proporcionales. Para una glutinosidad de la lengua escrita se puede tomar en cuenta (para el exterior de la palabra) la puntuación como resistencia (r), la que le sería indirectamente proporcional: $GL^{escrita} = \frac{kh}{r}$.

4.2.3 Las fronteras del sintagma y el carácter infinito del lenguaje

La aplicabilidad de las nociones de entropía y economía en el descubrimiento de fronteras

entre raíces y afijos, así como en la determinación de clíticos (o segmentos más gramaticales) es una invitación a examinar su posible pertinencia al descubrimiento de fronteras entre sintagmas. A continuación, se analiza cómo se podría estimar la glutinosidad entre sintagmas. En seguida se examinan los límites del cálculo de los factores de entropía y economía en cuanto a la estimación de la fuerza de adhesión entre sintagmas. Esto es importante para determinar hasta qué punto es pertinente el concepto de glutinosidad propuesto arriba al exterior del sintagma.

La cliticidad en las fronteras del sintagma

Al examinar los resultados de descubrimiento de clíticos del experimento descrito arriba, se observó que los valores de entropía y economía eran los más altos para los vocablos "más gramaticales". La relación de un clítico y un vocablo pleno se puede concebir en principio como un digrama (un par de segmentos). Pero lo cierto es que los segmentos gráficos pueden en efecto corresponder a unidades más complejas, por lo que el cálculo de índices de entropía y economía a partir de un digrama resultará necesariamente en una estimación más bien burda de la glutinosidad entre segmentos. De todas maneras, es natural que los índices más bajos de glutinosidad entre digramas sí hayan correspondido a fronteras entre sintagmas, por lo menos en el experimento llevado a cabo con el *CEMC*.

Para la lengua española esto era de esperarse, porque la estructura de los sintagmas del español es muy predecible; esto es, tienden a iniciarse con ciertos segmentos muy frecuentes de muy alta cliticidad (ciertos adverbios y pronombres personales en el sintagma verbal y preposiciones, artículos y ciertos adjetivos en sintagmas nominales), por lo que deparan menor

sorpresa (entropía) y no tienen una relación particularmente económica con los segmentos gráficos anteriores (que pertenecen a otros sintagmas).

En otras palabras, mientras que la entropía entre un clítico y un vocablo pleno tiende a ser comparativamente altísima (porque la incertidumbre de lo que sigue es considerable) y su relación económica tiende también a exhibir los valores máximos (porque el conjunto de lo que sigue es relativamente de gran tamaño), los valores de entropía y economía al inicio de un sintagma deben ser mucho menores, porque son pocas las cosas con las que tienden a iniciarse los sintagmas (un conjunto relativamente pequeño de "clíticos") y por lo general no causan gran sorpresa.

Nótese que esto no sería necesariamente igual si en lugar de calcular estos valores a partir de simples digramas (pares de palabras gráficas), se calcularan entre sintagmas completos (series de palabras gráficas con respecto a otras series de palabras gráficas), ya que la variedad de formas posibles al inicio de un sintagma sería mucho mayor. El problema es, sobre todo, que la mayoría de los sintagmas completos (aun en córpora enormes) tendría una frecuencia bajísima, casi siempre de uno (serían estructuras con un *hapax legomenon* incrustado). Además, no se ha mencionado nada de la posibilidad de estructuras anidadas, incluso oraciones, dentro de los sintagmas. Todo esto necesariamente afectaría el cálculo de cualquier índice.

Los límites del cálculo de los índices de entropía y economía

El problema principal de concebir la glutinosidad en términos de entropía y economía es que, mientras más larga sea la secuencia examinada, menor será la resolución de los índices

calculables, debido a la baja frecuencia de dichas secuencias y, por lo tanto, al bajo número de contextos en los que ocurran.

Después de un sintagma de una sola ocurrencia en un corpus dado sigue uno y solamente un sintagma (ya sea dentro o fuera de la oración o período), lo que significa que entre esos sintagmas hay una entropía mínima y un índice de economía bajísimo, aunque en la lengua misma quepa esperar una entropía o incertidumbre enorme al final de los sintagmas, porque cada signo constituido por un sintagma se puede combinar con una infinidad de sintagmas posibles para crear un nuevo signo en el nivel sintáctico (toda frontera entre sintagmas sería, en teoría, una articulación muy económica).

De hecho, en cuanto a la economía de signos se refiere, la noción de unos pocos signos perifrásticos de mayor frecuencia que se combinen con muchos otros de baja frecuencia se vuelve más difícil de manejar de manera cuantitativa y formal (incluso conceptual[12]) e implica de manera más directa que en el nivel morfológico cuestiones relativas a los fenómenos de significación y de cultura. Pero aun haciendo el intento de permanecer en la caracterización de la forma surgen otros problemas. Por ejemplo, queda la duda de qué otros conceptos de carácter lingüístico (aparte de la puntuación) pueden utilizarse como factores para determinar la fuerza de adhesión entre sintagmas. Debido al problema de la representatividad de los córpora (en los que se ha estimado que aproximadamente la mitad de los vocablos son *hapax legomena*), las nociones de economía y entropía (según se han estudiado en este trabajo) son sencillamente insuficientes para estimar la glutinosidad entre sintagmas. Por lo que no

[12] ¿Qué significa que un sintagma nominal sea un "presintagma" de uno verbal y vice-versa? o ¿qué significa que uno verbal sea una especie de "sufijo" del nominal y vice-versa?: ¿tenemos aquí jerarquías estables entre sintagmas de carácter sintáctico, de tal manera que un sintagma-sujeto sea una especie de "raíz" de un sintagma-predicado de carácter "sufijal"?

podemos por ahora hablar de este tipo de glutinosidad[13]. Sin embargo, creo que sí hay evidencia suficiente para definir una glutinosidad al interior del sintagma en términos de, cuando menos, entropía y economía.

4.2.4 Discurso y diacronía

En este subapartado se examina brevemente la manifestación de la glutinosidad en el tiempo. Primero se analiza el fenómeno discursivo, donde la fuerza de adhesión de los fragmentos léxicos entra en juego. En seguida, se considera la dimensión diacrónica, la cual involucra cuando menos dos estados de lengua.

Aunque las figuras 4.2 y 4.3 presentadas arriba (para ilustrar el pegamento entre las palabras) representan un segmento del discurso (`'De pronto se oyó un disparo sordo,...'`), hasta aquí solamente se ha tratado este fenómeno como fenómeno del sistema en sincronía. De esta manera, se ha establecido que cada segmento —ya sea un afijo, un clítico o una palabra gramatical— está asociado a cuando mucho dos valores de glutinosidad (uno a la derecha y el otro a la izquierda). Pero, como en esencia se trata de valores propios y, por ende, representativos para cada objeto, son valores sumamente abstractos.

Pero lo cierto es que esta fuerza debe variar en el discurso, cuando los objetos lingüísticos entran en juego en el proceso de significación. Es decir, la dimensión temporal ausente del sistema sincrónico también es de pertinencia a la energía glutinosa. Así, aunque los afijos

[13]Esto no excluye la posibilidad de utilizar las marcas de puntuación como evidencia cuantitativa de fronteras de sintagmas. De hecho, en este experimento la puntuación por sí misma resultó muy confiable en el descubrimiento de este tipo de fronteras. Sin embargo, no podemos considerar este fenómeno como un universal del lenguaje, sencillamente porque la mayoría de las lenguas no se escriben y, finalmente, no hablamos con signos de puntuación.

se adhieran a una clase abierta de segmentos, es natural que no todos se peguen a todos los segmentos posibles con la misma fuerza: mientras que *~ando* se adhiere sobre todo a raíces de verbos, no lo hará con la misma fuerza a otros tipos de segmentos: **comprarando.* **posibleando.* También los clíticos tienden a ocurrir en determinados contextos, de tal manera que en español, por ejemplo, la glutinosidad entre un pronombre proclítico y un verbo ('se duerme') será naturalmente mayor (aun cuando el verbo particular no acepte muy bien al proclítico en cuestión: 'lo muere') que aquella entre un artículo y un pronombre proclítico ('los me') o un proclítico y un sustantivo ('te contrariedad').

Pero como hemos visto, en una muestra discursiva como el *CEMC* bien puede ocurrir cualquier cosa, por error o por lo que sea. Y esto se debe reflejar en la fuerza glutinosa del discurso. Es decir, si una combinación de signos ocurre, hay que tener las herramientas para analizarla. Incluso, si nos incumbe la tarea de juzgar la "gramaticalidad" (ahora sí en su sentido de aceptabilidad), esto se debe reflejar en las cantidades de glutinosidad involucradas. Éstas, claro está, pueden servir solamente de base al inevitable examen cualitativo-introspectivo, pero con la ventaja de basarse en datos verdaderamente empíricos.

Por otra parte, también es posible seguirle la pista al fenómeno de la glutinosidad entre dos o más estados de lengua, es decir, en la diacronía. Si, como Sapir señaló, la glutinosidad crece y decrece a través de los milenios, esto debe en teoría reflejarse en las estimaciones cuantitativas de esta energía. Desde luego, esto implica grandes problemas por lo menos en cuanto a la codificación de las formas en distintos estados de lengua. Es decir, las variaciones ortográficas a lo largo del tiempo (la necesidad de recurrir a la paleografía) y el desgaste fonológico típico de las formas que dejan de ser plenas, implican un laborioso trabajo por

parte del analista para establecer la identidad entre objetos de un estado y otro. Por ejemplo, en español las variadísimas formas del verbo 'haber': *aber*, *~emos*, *~ían*, etc., o el enorme desgaste de *vuestra merced* a *usted*.

Pero todas estas complicaciones no quitan que se pueda estimar y comparar la glutinosidad en diferentes estados de las lenguas. De hecho, si nos incumbe el estudio del fenómeno de "gramaticalización" (en el sentido de fosilización de segmentos), las formas que muestren un incremento de glutinosidad a lo largo del tiempo (más allá de cualquier corazonada o impresión) serán los mejores candidatos para ilustrar dicho fenómeno. También aquí, está claro que el toque definitivo en cualquier análisis será proporcionado por el examen cualitativo del analista, pero basado en criterios formales.

4.3 Teoría de la medición: hacia una glutinometría formal

En este apartado se examina el problema de construir un esquema de medición en general —cosa fundamental a todo tipo de ciencias— para luego proponer un esquema de medición de la glutinosidad, es decir, una glutinometría. En primer lugar, se presentan algunas generalidades de la teoría de la medición. Luego, se determinan las dimensiones pertinentes a la medición de la glutinosidad. Después, se examinan las posibles unidades de medición en este esquema. Más tarde se presentan los axiomas necesarios para una teoría glutinométrica. Por último, se plantea el problema inevitable del error en todo proceso de medición.

4.3.1 Generalidades de la medición

En esta sección se presentan algunas generalidades de la teoría de la medición que nos permitan acercarnos a la construcción de un esquema glutinométrico. En concreto, se examinan los conceptos de cuantificación, cantidad o magnitud, dimensión, unidad y escala de medición. Como se verá, estos conceptos son cruciales para construir cualquier esquema de medición.

Cada vez que se le asignan números a los rasgos de un fenómeno, mediante algún proceso de observación, se está midiendo algo. Es decir, la medición es el proceso que consiste en asociar números a objetos, fenómenos o cantidades físicas. La teoría de la medición se encarga del estudio de cómo se asignan dichos números a fenómenos y magnitudes físicas de diversas índoles. Hay tantos tipos de medición como tipos de propiedades y técnicas de medición, pero en esencia, medir es contar o comparar.

Pero antes de que se pueda llevar a cabo un proceso de medición, es necesario entender lo que es el proceso de *cuantificación* numérica que precede a toda medición[14]. Este proceso es una operación conceptual y consiste en asociar conceptos a variables numéricas[15]. Según Bunge, hay varias maneras de asignar un número a un concepto: como nombre, como probabilidad o proporción, como cantidad de objetos (cardinalidad) o como la cantidad de alguna propiedad del concepto[16].

---

[14]Bunge, *Philosophy of Science II: From Explanation to Justification* [27], Transaction Publishers, New Brunswick, 1998, pp. 217-228.

[15]También puede entenderse como un refinamiento de formulas cualitativas.

[16]El tipo de asociación entre números y conceptos, depende sobre todo del tipo de objeto en cuestión (ya que éste puede ser un individuo —'x', 'c'—, una clase —'viviente', 'metal'—, una relación sea o no comparativa —'$\leqq$', '$\in$'— o una cantidad —población, longitud); véase Bunge, *Philosophy of Science I: From Problem to Theory* [26], Transaction, New Brunswick, 1998, p. 67.

Una *cantidad, magnitud* o —como dice Bunge— *predicado métrico* o *functor numérico*[17] es un concepto complejo que puede concebirse como una agrupación de objetos variables x, variables numéricas y y una función de los objetos a las variables $L(x) = y$ (que puede representar, por ejemplo, una cantidad lineal: 'la longitud de x es y'). Este concepto cuantitativo es una manera precisa de representar la propiedad de uno o varios objetos dados.

Un aspecto muy importante, previo al proceso de medición, es lo que se conoce como el *mesurando*, es decir, la cantidad empírica que ha de medirse[18]. Así, la medición se lleva a cabo al comparar dicho mesurando con alguna cantidad estándar del mismo tipo. La cantidad predeterminada o estándar se establece arbitrariamente o mediante la referencia a alguna constante universal. En principio, el mesurando puede ser cualquier cosa, pero en esencia es un valor numérico particular de un concepto cuantitativo (es decir, el valor de una magnitud, obtenido empíricamente). Con base en esto, se distinguen por lo menos tres conceptos en la medición:

Mesurando o propiedad concebida en grados: Para una propiedad que ha de ser medida objetivamente, se presuponen grados de valor. Los grados, cantidades o intensidades de esta propiedad se representan mediante $\dot{r}$. El conjunto de estas cantidades se representa $\dot{R}$. Se asume que $\dot{r}$ es una característica de la realidad, que es independiente de toda medición y conceptualización, y que bajo ciertas circunstancias es perceptible.

Valor medido (del mesurando): El valor del mesurando es la cantidad de grados $\dot{r}$ en el objeto que ha de ser medido. Se representa mediante la función $m(\dot{r})$. El conjunto de estos valores se representa $\{m(\dot{r})\}$. El valor de la función $m(\dot{r})$ resulta de la observación, es decir, de un proceso empírico de medición, por lo que se dice que es una estimación de $\dot{r}$.

Valor numérico: Los valores numéricos de una magnitud que representa una propiedad se designan mediante r. El conjunto de estos valores es R. Estos valores son parte del aparato teórico.

---

[17] Bunge, *Scientific Research II. The Search for Truth* [25], Springer-Verlag, Berlin/Heidelberg, 1967, pp. 198-199.

[18] *Ibid.* [25], p. 206.

Tabla 4.1: El mesurando y sus estimaciones

			ejemplos	
valores objeto[a]	**símbolo**	**nivel**	**longitud**	**glutinosidad**
mesurando (grados de la propiedad)	$\dot{r}$	realidad	←longitud→	$s_1 \mapsto s_2 \Leftrightarrow s_3 \leftrightarrow s_4$
valor estimado de la propiedad	$m(\dot{r})$	experiencia	(100 ± 0.1) cm	resultados de un programa glutinómetro
valor numérico de la propiedad	r	teoría	100 cm	valores idealizados y asociados a los tipos s_1, s_2, s_3, s_4

[a]Tabla basada en la de Bunge, *op. cit.* [25] 1967, p. 207.

En la tabla 4.1 se resumen las características de estos objetos. Las dos últimas columnas representan ejemplos: los valores involucrados en el proceso de medir longitudes y los de una glutinometría posible. Así, al asumir una fuerza de atracción entre elementos del lenguaje medible en grados de algún tipo, es importante distinguir esas fuerzas de los valores que se puedan calcular mediante algún programa computacional (siempre sujetos al método, la muestra y los errores) que, además, de ninguna manera pueden confundirse con los valores que se puedan asumir como pertinentes a las formas de una lengua (que serían objetos teóricos — resultados y herramientas, no sólo de investigaciones cuantitativas, sino también de trabajos cualitativos).

Lo importante de esta concepción es señalar que hay una discrepancia entre los valores reales $\dot{R}$, por un lado, y los valores medidos $\{m(\dot{r})\}$ (y, por lo tanto, los numéricos R), por el otro. Y casi siempre esta discrepancia es desconocida y rara vez se puede determinar. Esto lleva al problema del error, que se examina en la última sección de este apartado.

En cuanto al proceso mismo de medir, si bien se puede llevar a cabo mediante la pura percepción sensorial de quien *mide* (en cuyo caso se trata más de apreciaciones subjetivas que de mediciones propiamente), se realiza generalmente con instrumentos de algún tipo.

que pueden ir del simple escalímetro para medir longitudes a dispositivos muy complejos, capaces de medir fenómenos fuera del alcance de los sentidos. Así, al leer un texto, el ojo educado podrá percibir, auxiliado de las pautas proporcionadas por la cultura (tales como los espacios y otros signos de puntuación entre palabras gráficas), qué segmentos son los que le dan estructura al texto (y cuáles lo hacen más). Sin embargo, un glutinómetro podría medir las asociaciones gramaticales entre los segmentos del texto de manera mucho más fina.

Existen varios tipos de procesos de medición, pero el más básico es el simple acto de contar objetos observables. Una de las suposiciones importantes es que este proceso no afecta a los objetos observados ni a los medios mediante los cuales se detectan cambios causados por éste en los objetos en cuestión[19]. Contar significa simplemente establecer una correspondencia de uno a uno entre un conjunto de objetos (en cualquier orden) y un subconjunto de los enteros positivos (0, 1, 2, 3, etc.).

Dependiendo del tipo de procedimiento para contar, éste puede hacerse directamente o mediante procesos basados en conceptos teóricos muy elaborados (por ejemplo, para contar moléculas). El requisito principal para que una colección de objetos o hechos se pueda contar es que contenga miembros empíricamente distintos, es decir, que se distingan entre sí y estén separados: que sean discretos y que esto se pueda discernir directamente mediante los sentidos (directamente contables) o con la ayuda de herramientas especiales (indirectamente contables), incluyendo el muestreo y los aparatos conceptuales teóricos aplicables a la colección de objetos[20]. Una glutinometría presupone que los ladrillos lingüísticos, que son

[19]Bunge examina con detalle los problemas del proceso de contar, *ibid.* [25], pp. 213-218.

[20]*Ibid.* [25], p. 214.

los morfemas, son empíricamente distintos. Si la cultura de la cual proviene el corpus no proporciona las marcas de puntuación que delimiten los ladrillos aceptados culturalmente (palabras gráficas rodeadas de espacios), éstos son empíricamente accesibles mediante sus relaciones económicas y de información con el resto del corpus. Las marcas culturales los hacen hasta cierto punto directamente contables. Un glutinómetro, por otra parte, serviría para contarlos indirectamente y de manera más fina, de tal manera que lo que no se ve a simple vista se puede medir empíricamente.

Otro aspecto importante del proceso de medición se refiere en especial a la contabilidad de *grados* de propiedades continuas (no discretas). Para medir los grados o la *intensidad* de una propiedad, las magnitudes se pueden comparar a una cantidad estándar o *unidad* de medición predeterminada arbitrariamente o mediante la referencia a alguna constante universal. Se trata de averiguar empíricamente cuántas unidades caben en una cantidad dada. De esta manera, si la cuantificación de una propiedad consiste en asociarle variables numéricas, la medición es la contraparte empírica de la cuantificación y consiste en interpretar las partes equivalentes a las unidades como una imagen más o menos cercana a los grados de la propiedad a medirse. Al llevar a cabo un acto de medición, el observador percibe una marca m^* en el instrumento de medición, la que interpreta como la imagen de un número $m(\dot{g})$ (una estimación del valor real g, que casi siempre se desconoce). Entonces, se asume que $m(\dot{g})$ es una fracción cercana a g (véase la tabla 4.2).

Todo esto nos permite ver que la medición es en realidad un proceso muy complejo. Además de unidades, los sistemas de medidas están constituidos por *escalas*. Una escala es el intervalo con que se representan los grados de una propiedad y está constituida por repre-

Tabla 4.2: Correspondencia entre los grados de una propiedad y sus equivalentes instrumentales con respecto a los números

sistema[a] **concreto**	**propiedad física** (grados objetivos)	**propiedad conceptual** (números reales[b])	**lectura instrumental**	**medida de la propiedad** (números racionales)
□□	$\dot{g}_1$	g_1	m_1^*	$m(\dot{g}_1) = [g_1]^c$
□□□□□	$\dot{g}_2$	g_2	m_2^*	$m(\dot{g}_2) = [g_2]$
□□□□	$\dot{g}_3$	g_3	m_3^*	$m(\dot{g}_3) = [g_3]$

[a]Tabla basada en la de Bunge, *op. cit.* [25] 1967, p. 220.
[b]Números desconocidos.
[c]Valor cercano a g_1.

sentaciones ordenadas y espaciadas de las unidades. Puede ser material, como un instrumento (por ej. un escalímetro, que consta de un conjunto ordenado de marcas), y conceptual, como un intervalo numérico imaginado (los números reales entre [0,1]).

Entonces, si se empieza por asumir un sistema de relaciones de los hechos que corresponde a uno imaginado (cuantificación), luego se asume la existencia de otro sistema escalar de unidades tanto material como conceptual que participan en el proceso de medición. Es decir, se empieza por asumir la existencia de un sistema real de relaciones $\dot{\mathbb{R}} = \langle \dot{R}, \dot{\leqq} \rangle$, en donde $\dot{R} = \{\dot{r}\}$ es el conjunto de grados de una propiedad física y '$\dot{\leqq}$' es la relación concreta de ordenamiento 'es más largo o igual de largo que', o 'es más afijo o igual de afijo que'. Una vez asumido este sistema, entonces se puede imaginar un sistema conceptual de relaciones $\mathbb{R} = \langle R, \leqq \rangle$, compuesto del subconjunto R de números reales (un segmento del continuo de números reales) y la relación aritmética $\leqq$, 'es menor o igual a'. Entonces, para cuantificar la propiedad que se desea medir, se asume un *isomorfismo* entre $\dot{\mathbb{R}}$ y $\mathbb{R}$ (véase la tabla 4.3).

De manera similar, $\mathbf{M}^* = \langle M^*, \dot{\leqq} \rangle$ representa al sistema real de relaciones que consiste en el conjunto $M^* = \{m^*\}$ de marcas en un escalímetro (o cualquier instrumento de medición) ordenadas mediante la relación física $\dot{\leqq}$ ('a la izquierda o igual que'). Además, $\mathbf{M} = \langle M, \leqq \rangle$ es

Tabla 4.3: Relación entre cuantificación y medición con respecto a los números

hechos[a]		**ideas**
	cuantificación	
$\dot{\mathbb{R}} = \langle \dot{R}, \dot{\leqq} \rangle$	$\xrightarrow[\text{correspondencia}]{}$	$\mathbb{R} = \langle R, \leqq \rangle$ (valores numéricos reales)
	medición	↑correspondencia parcial
$\mathbf{M}^* = \langle M^*, \dot{\leqq} \rangle$	$\xrightarrow[\text{correspondencia parcial}]{}$	$\mathbf{M} = \langle M, \leqq \rangle$ (valores medidos irracionales))

[a]Tabla basada en Bunge, *op. cit.* [25] 1967. p. 221.

el conjunto correspondiente de números racionales (un subconjunto de los números reales). Entonces, el resultado de una serie de mediciones de una propiedad dada se lleva a cabo estableciendo una correspondencia parcial entre el sistema empírico de relaciones $\mathbf{M}^*$ y el sistema conceptual $\mathbf{M}$, que es —como se dijo arriba— una imagen parcial de $\mathbb{R}$, que a su vez es una imagen de $\dot{\mathbb{R}}$. Si se elimina cualquiera de estos sistemas $\dot{\mathbb{R}}$, $\mathbb{R}$, $\mathbf{M}^*$, $\mathbf{M}$ o se ignora cualquiera de las correspondencias, no se puede decir que haya un acto de medición[21].

Las escalas son construcciones del reino de las ideas, por lo que están constituidas por los sistemas de la derecha en la tabla 4.3: la escala conceptual sería $\mathbb{R}$ y la escala material $\mathbf{M}^*$. Hay varias cosas que considerar de su naturaleza. Primero, las escalas necesitan de un origen: dónde se coloca el cero, lo dicta la magnitud que se quiere medir y los conocimientos que se tengan de ella. El origen puede ser convencional (como en la temperatura) o absoluto (como en la edad, que se cuenta en el intervalo $[0,\infty]$). En los primeros el cero puede ir en cualquier lado, en los segundos casi siempre hay una propiedad que se desvanece[22]. Segundo, hay varios tipos de escalas. Pueden ser ordinales o topológicas (cuando están constituidas de marcas sin relación entre sí, por ej., dispuestas caóticamente) o métricas (cuando existe un espaciamiento sistemático, no necesariamente uniforme). Y, tercero, adoptar una escala

[21] *Ibid.* [25], p. 221.

[22] *Ibid.* [25], p. 223.

y no otra es una cuestión práctica y no de verdad[23]. En el marco de una glutinometría, se puede concebir una escala métrica cuyo origen estaría vinculado al desvanecimiento de la glutinosidad: un valor de cero correspondería al origen absoluto de un intervalo $[0.\infty]$ o, en una escala normalizada, [0,1].

Pero volviendo a la discusión sobre los grados de las propiedades susceptibles de ser medidas, toda escala métrica requiere de una unidad o intervalo básico, tanto para la parte conceptual como la material. Como se dijo arriba, el acto de medir consiste en determinar empíricamente cuántas unidades caben en una magnitud. La manera usual de expresar magnitudes es:

$$P(x, s) = ru$$

donde $P(x, s)$ es la propiedad del objeto x a medirse en la escala s (P no es un número). u es la unidad (tampoco es un número) para medir P y r es el valor conceptual de la magnitud.

Las magnitudes corresponden a *dimensiones*. De hecho, puede haber diferentes magnitudes para cada dimensión: es decir, cada dimensión puede tener una familia de magnitudes. Además, las magnitudes pertenecientes a una misma familia se pueden medir todas con las mismas unidades. Así, el concepto de longitud es una dimensión que representa a varias magnitudes tales como la altura y la distancia. Similarmente, la afijalidad y la cliticidad pertenecen a la misma familia que bien podemos denominar glutinosidad.

Las unidades se pueden determinar solamente después de un análisis de la dimensión que van a medir. Este análisis no es necesario cuando se trata de una magnitud simple o fundamental, como el tiempo o la longitud. Pero la mayoría de las magnitudes son derivadas

[23] *Ibid.* [25], p. 222.

o complejas, es decir, se describen mediante fórmulas que las relacionan con las magnitudes fundamentales. Por ejemplo, la velocidad se define $v = d/t$ (distancia sobre tiempo). Todas estas fórmulas pueden ser de naturaleza tanto empírica como teórica, pero siempre pertenecen o presuponen alguna teoría[24]. Así, la dimensión glutinosidad —que, como hemos visto, podría derivarse a partir tanto de una medida de información como de una de economía— es medible mediante la fórmula $GL = kh$ (economía por entropía) que es de naturaleza a la vez empírica —al medir una propiedad de objetos de la realidad— y teórica —porque presupone la teoría de la información y la noción de economía en el lenguaje. A continuación, aparece un análisis más detallado de la glutinosidad y las dimensiones a partir de las cuales se puede derivar. Después, una vez hecho este análisis, se examinan finalmente las unidades de la glutinometría.

4.3.2 Las magnitudes y las dimensiones de la medición

En este apartado se presentan algunos principios generales del concepto de dimensión y se exploran las dimensiones pertinentes a cualquier método de medición de la glutinosidad. Es decir, después de definir algunos conceptos básicos, una por una de las estrategias para segmentar discurso se analizan con el objeto de determinar las magnitudes (y dimensiones fundamentales) más aptas para medir esta energía de adhesión.

Como se estableció arriba, el concepto de cantidad o magnitud es de naturaleza cuantitativa. Es una manera precisa de representar mediante valores numéricos una propiedad o atributo de los objetos estudiados. Una dimensión, por otra parte, no es la magnitud de una propiedad, sino la propiedad misma. Cada propiedad o atributo de un objeto que sea sus-

[24] Bunge, *ibid.* [25], p. 225.

ceptible de medirse (y obtener así magnitudes representativas de dicha propiedad) constituye una dimensión. En ese sentido las magnitudes *miden* dimensiones.

También se dijo arriba que hay familias de magnitudes para cada dimensión y que las magnitudes pertenecientes a una misma familia se pueden medir todas con las mismas unidades. Entonces, así como la dimensión longitud L representa a las magnitudes distancia, altura, etc., la dimensión entropía I representa a magnitudes tales como la incertidumbre, sorpresa, información, etc., y la dimensión G (en cierta manera, gramaticalidad) representa a las magnitudes de afijalidad, cliticidad y glutinosidad.

Puesto que no todas las dimensiones son independientes, algunas se pueden expresar en términos de otras, es decir, como función de otras dimensiones, cosa fundamental al análisis dimensional y la existencia de sistemas de unidades. En otras palabras, algunas dimensiones, en su calidad de atributos de los objetos, se derivan de otras dimensiones, llamadas fundamentales. Por ejemplo, así como la velocidad y la aceleración se derivan del tiempo y del espacio, la cantidad de economía se expresa en términos de signos (que aquí designaremos S) que resultan de otras cantidades de signos, es decir, la economía mide el promedio del número de signos nuevos del nivel superior, creados mediante la combinación de unos pocos signos muy frecuentes y muchos otros poco frecuentes del nivel anterior. Así, la cantidad de economía es en realidad una magnitud derivada de dos valores (primarios) cuya dimensión es también el número de signos: $\frac{S}{S}$. En la tabla 4.4 se resumen las dimensiones pertinentes a la glutinosidad.

Pero antes de continuar examinando las magnitudes y dimensiones de la glutinosidad, en los próximos párrafos se contrastarán las dimensiones de entropía (I) y economía de signos ($\frac{S}{S}$)

Tabla 4.4: Dimensiones de la Glutinometría

dimensión	tipo	descripción
S	primaria	cantidad de signos; es una familia de varias magnitudes: alternantes, acompañantes, topogramas, etc.
I	primaria	cantidad de sorpresa o información
$G = \frac{S}{S} \times I$	derivada	cantidad de información en signos económicos

con las otras magnitudes provenientes de las otras estrategias de segmentación del discurso para argüir que las primeras están en la esencia del fenómeno, mientras que las segundas son meros efectos de éste.

Las dimensiones como causas y como efectos

En esta subsección se comparan aquellas medidas que parecen ser apenas resultado o consecuencia del fenómeno de glutinosidad con aquellas que parecen definirlo. En concreto, se examinan brevemente los factores de frecuencia absoluta y número de cuadros (y otras combinaciones), cuya omisión en la estimación de esta fuerza de adhesión no ha estado del todo justificada.

La frecuencia absoluta de las formas pertenece a la dimensión S, es decir, se trata en esencia de cuentas de signos. Esta magnitud tiene, en los estudios del lenguaje, un lugar privilegiado. Se toma como evidencia de las más variadas cosas. Y, en efecto, se trata de un indicio de qué tan importantes y no marcadas son las formas examinadas. Pero por lo general no es una medida muy fina. Por eso se llegan incluso a aplicar procedimientos que la "corrigen" para reflejar, por ejemplo, su dispersión en diferentes géneros[25]. De hecho, la frecuencia se puede interpretar incluso como una medida de entropía (las formas más frecuentes son necesariamente las menos informativas) o economía (las formas más frecuentes

[25] Véase discusión sobre Lara y Ham, art. cit. [88] 1974, a partir de la página 38 del capítulo anterior.

se están combinando con las otras para formar nuevos signos). Pero si se trata de medir dichos fenómenos, podemos recurrir a los procedimientos ya conocidos, que dan resultados más finos.

Por otra parte, el número de cuadros (y otras combinaciones) también es, en esencia, la cuenta de formas pero más complejas. De alguna manera, también se trata de una medida de desorganización o entropía. Pocos cuadros (u otras estructuras combinatorias) implican un corpus demasiado pequeño o una combinación de signos poco organizada. Por otra parte, muchos cuadros entre dos segmentos implica mucha información, mucha entropía, pero ningún indicio de la dirección en que los segmentos "sorprenden" (cuál de los dos es el signo más gramatical y cuál el de contenido). Además, esta medida también puede interpretarse como una de economía, porque muchos cuadros implican muchos signos del nivel siguiente. Pero por muchos cuadros que se detecten en una segmentación, no hay ninguna evidencia de que sean producto de unos pocos (pero frecuentes) segmentos combinados con otros muchos de baja frecuencia. Es decir, tenemos un indicio de signos nuevos de un nivel (evidencia de economía), pero no tenemos acceso a su estructura interna.

Indudablemente, estas dos medidas sí son evidencia de las posibles segmentaciones de un corpus, principalmente porque llevan implícitos tanto los criterios de entropía como los de economía. Por eso funcionan. Sin embargo, ambas están sujetas a fluctuaciones importantes (que dependen en mucho de la representatividad del corpus, su estructura y los siempre inevitables errores). Esas fluctuaciones parecen ser ruido agregado a los fenómenos de equiprobabilidad y economía, como si estos últimos estuvieran en el corazón mismo del corpus (como si fueran su causa misma), mientras que las meras frecuencias de las formas (simples y complejas) aparecen apenas como efectos de esos latidos, cuyos mejores indicios

son las medidas de información y economía de signos (los cuales, después de todo, también son evidencia, no el fenómeno mismo).

Otra de las medidas que es necesario examinar es la contabilidad de signos de puntuación. Esta medida, que también pertenece a la dimensión S, parece medir más un efecto que una causa de la fuerza de glutinosidad: la puntuación se coloca donde hay menos de esta pegajosidad lingüística. Sin embargo, en los experimentos del capítulo anterior resultó ser un indicio excelente de las fronteras entre sintagmas y, por ende, de los segmentos más clíticos (la puntuación tiende a no ocurrir entre los clíticos y las palabras a las que se adhieren). Por eso, vale la pena tomarla en cuenta más para afinar el cálculo de las cantidades de glutinosidad, que como factor definitorio de esta fuerza. Así, al cabo de las siguientes subsecciones dedicadas a la entropía y la economía, se examina la cuestión de la puntuación, como magnitud pertinente en la segmentación del discurso escrito.

El contenido de información

En esta subsección se establece la necesidad de utilizar una medida de la información transmitida entre las unidades del discurso para determinar la glutinosidad formal. Esto es pertinente por la presunción de que en todas las lenguas las unidades menos informativas son las más gramaticales.

Como se mencionó en el primer capítulo, ya Greenberg había notado que enunciar uno de los miembros de una clase de morfemas que funcionan como raíces implica un contenido informativo mayor (que de morfemas que funcionan como afijos) en ambos sentidos del con-

cepto de información, el técnico y el no técnico[26]. Creo que esto lo hemos comprobado en los dos capítulos anteriores. Sin embargo, es muy claro que el contenido de información —en el sentido técnico— no es lo mismo que el significado o contenido semántico de un signo. Por eso, para determinar la condición definitiva de morfema que pueda tener un segmento, sigue siendo importante tomar en cuenta su significado.

Desde el punto de vista cuantitativo, el problema del significado propio implica muchas variables y requiere la intervención intensiva y constante del analista, quien es finalmente responsable del trabajo cualitativo. Por eso, resulta especialmente ingenioso que una medida de desorganización indique la cantidad de información transmitida, al ser las estructuras más organizadas las menos informativas y las más caóticas (menos probables) las más informativas. Así, aunque signifique muchas cosas un término tan frecuente (y repartido en todos los géneros del *CEMC*) como 'trabajo', éste transmite poquísima información porque casi podemos esperar que ocurra en el siguiente párrafo sin que esto nos dé indicios sobre la información contenida en ese párrafo. Mientras que un término menos común como 'chamba' (que contiene apenas uno de los significados de 'trabajo' en un registro coloquial) nos daría más información sobre el párrafo siguiente, si supiéramos de antemano que allí ocurre. De igual manera, si por un lado supiéramos que la palabra 'de' ocurre en el párrafo siguiente, no tendríamos mucha información sobre lo que allí se dice, pero si por el otro supiéramos que aparece 'braquistiquio', este término nos estaría dando muchísima información sobre ese párrafo, aun sin conocer su significado.

Pero si bien es cierto que el significado no es lo mismo que información o entropía (porque

[26]Greenberg, *op. cit.* [58] 1957, p. 91.

se trata de una propiedad entre signos y no de una contabilidad de significados al interior de éstos), también es cierto que el primero es inseparable de la forma lingüística. Digamos que deja su huella en la equiprobabilidad de las formas: por donde transita el significado, queda una huella de su paso. Y esa pista no queda propiamente en las formas mismas, sino en sus relaciones con las demás formas. Por eso, en parte, es lógico que la glutinosidad —al ser una energía lingüística inevitablemente relacionada al significado de las formas— pueda expresarse mediante una medida de información: la glutinosidad es información transmitida.

Además de Greenberg, también Harris había entendido la importancia de este fenómeno en la lingüística. Las cuentas de fonemas a la izquierda y a la derecha que estudió para segmentar el discurso son aproximaciones a la entropía: hay más caos donde las cuentas son mayores, cosa que implica un corte entre segmentos. Pero si bien es cierto que esas cuentas no miden la equiprobabilidad de los fonemas posibles, también es cierto que Harris mismo después reconoció la importancia de ese fenómeno en el lenguaje[27].

Un índice de economía

En esta sección se establece la necesidad de contar con una medida del ahorro o economía de signos en el cálculo de una glutinosidad formal. La idea es que esta propiedad tampoco es exclusiva del español y que, de hecho, debe jugar un papel importante en otras lenguas.

Dentro y fuera de la lingüística, el concepto de economía tiene varios sentidos. En esta investigación la acepción que se ha utilizado se refiere a la combinación de signos de un nivel para crear signos de número potencialmente infinito del siguiente nivel. El índice de

[27] Harris, *op. cit.* [66] 1991, pp. 22-23, 30-33.

economía pensado por de Kock es una medida de la relación entre los signos del nivel inferior con respecto a los signos atestiguados del nivel superior: mientras menor sea el número de signos del primero y mayor sea el de los del segundo, más economía de signos habrá en la lengua examinada.

En cierta manera, el procedimiento de Harris de contar fonemas anteriores y posteriores también resulta en una medida de economía. Las cuentas mayores de fonemas adyacentes a un segmento de vocablo corresponden a la mayor numerosidad de formas a las que dichos fonemas pertenecen y que ocurren en el corpus junto a dicho segmento (esto es, los fonemas son un indicador de cuántas formas acompañan a dicho segmento). Además, si estas cuentas de fonemas se ven como las cuentas de los signos que acompañan al segmento examinado, entonces el procedimiento de Harris se parece al caso de medida de economía en que el número de signos que alteran con este segmento es 1, cosa casi siempre falsa, diría seguramente de Kock, para quien esa relación entre unos pocos signos muy frecuentes y unos muchos de baja frecuencia para formar nuevos signos del nivel superior es crucial.

Pero no se puede asumir que todas las lenguas recurran exactamente de igual manera a esta estrategia de ahorro de signos, tan común en lenguas indoeuropeas. El fenómeno de economía está sin duda presente de varias maneras en otras estrategias, también susceptibles de medirse cuantitativamente. Tómese, por ejemplo, el aprovechamiento de posibilidades combinatorias de cierto número de signos en secuencias de una longitud dada. Altmann observa que:

> Die Permutierbarkeit der Wörter im Satz ist in jeder Sprache mehr oder weniger eingeschränkt: daher kann man das Verhältnis der erlaubten un der theoretisch möglichen Wortpermutationen als ein Maß der Strenge der Wortstellung betrachten. Der Anteil der Phonempermutationen, die sinnvolle Wörter ergeben,

deutet das Maß der phonematischen Wortbildung an. Je größer dieser Anteil. desto größer die Wortbildungsökonomie[28].

De esta manera, al contar el número de combinaciones de signos atestiguadas en un corpus y compararlo con el número matemáticamente posible de combinaciones, tenemos otro índice de economía no necesariamente ligado a los fenómenos de afijación y cliticización. Aquí no podemos sino especular que aquellos grupos de signos aprovechados en el máximo número de secuencias, tendrán un carácter más gramatical que el resto, es decir, es de esperarse que exhibirán entre ellos un alto grado de glutinosidad. Lo que es claro es que. si el lenguaje es por naturaleza económico, tiene que haber un procedimiento para medir esa manera de ser económico y cabe esperar que ese ahorro formal esté en relación directa con la fuerza de asociación entre palabras.

Lo importante es resaltar que la economía de signos favorece a los hablantes al ahorrarles tiempo de aprendizaje del inventario de signos y memoria para recordarlos. Podemos suponer, entonces, que las lenguas con un inventario pequeño recurrirán a menos estrategias económicas, pero no podemos concluir que no haya en ellas algo de economía. En ese sentido siempre habrá economía que pueda medirse.

La puntuación como un tipo de resistencia

En esta sección se argumenta a favor de la inclusión de alguna cuenta de signos de puntuación como dimensión pertinente en el cálculo de la glutinosidad, especialmente entre sintagmas. Es importante tomar en cuenta esta cuestión, principalmente porque los córpora de

[28] Véase Altmann, *Statistik für Linguisten* [6], Wissenschaftlicher Verlag, Trier. 1995. p. 31.

cualquier lengua, por más que se trate de lengua hablada, llevan inevitablemente las marcas de la lengua escrita.

La relación entre lengua hablada y lengua escrita es muy compleja. De hecho, en este espacio no se le puede hacer justicia a semejante problema, sobre todo si se trata de examinar el fenómeno de las lenguas en general y no nada más del español. De todas maneras, vale la pena mencionar algunas generalidades.

Por ejemplo, parte de este problema es el hecho de que ni siquiera hay consenso en cuanto a la naturaleza de la dualidad oral-escrito. Existen por lo menos cuatro actitudes que pueden tomar los lingüistas[29]:

1. identificar a la lengua escrita con *la* lengua (actitud en gran parte subyacente a la gramática tradicional),
2. identificar a la lengua oral con *la* lengua: la lengua escrita es apenas una representación deformada de ella (fonocentrismo),
3. considerar que la lengua es fundamentalmente de naturaleza hablada, pero que lo escrito la representa de manera "bastante fluida" (fonografismo), y
4. considerar que la lengua existe bajo dos formas entre las cuales la lingüística no postula ni jerarquía ni dependencia (postura autonomista).

Pero estas actitudes no necesariamente están compitiendo entre sí cuando se trata de reflejar la realidad tipológica de las lenguas naturales. Por ejemplo, la postura 2 es inevitable al estudiar las lenguas sin escritura. Pero, por otra parte, la 4 parece describir más bien situaciones extremas, como la del japonés del siglo XIX, que padecía de diglosia extrema entre la lengua hablada y la escrita[30]. Lo cierto es que las lenguas escritas fueron, al principio,

[29]Anis, "¿Una grafemática autónoma?" [9], pp. 271-272, en Catach, ed., *Hacia una teoría de la lengua escrita* [31], Gedisa, Madrid, 1991.

[30]Véase Coulmas, "Superación de la diglosia: acercamiento del japonés escrito y hablado en el siglo XIX" [45], en Catach, ed., *op. cit.* [31] 1996, pp. 242-256.

lenguas solamente habladas y que una lengua sin escritura sencillamente no es lo mismo que la misma lengua con escritura. Catach intenta capturar esto al proponer que "todo lenguaje L provisto de un oral A y de un sistema de escritura desarrollado B" se convierte en L'[31]. cosa que simboliza mediante la expresión $A \times B = L > L'$, donde A es un lenguaje hablado. B es un sistema de escritura y L' es la combinación de ambos.

Además, la simple concepción del lenguaje hablado como diferente al lenguaje escrito implica la existencia no de dos, sino de por lo menos cuatro tipos de lenguaje[32]:

hablado: discurso directamente codificado por un hablante,
escrito: discurso directamente codificado por el escritor,
oralizado: discurso escrito que se habla (por ej. la lectura en voz alta), y
transcrito: discurso hablado que se escribe (por ej. la transcripción de entrevistas).

En este contexto, no está de más recordar que la puntuación es un fenómeno de lengua escrita y que, por lo tanto, tomarla en cuenta en el cálculo de la glutinosidad, le daría a esta última un fuerte matiz textual, en el sentido de consignar un fenómeno no necesariamente oral. Pero también es cierto que no es extraño que se utilice la puntuación y otros símbolos como un tipo de notación para representar pausas, particularidades prosódicas, hechos extralingüísticos, etc., así como para evitar ambigüedades en la transcripción al papel de una lengua hablada (lenguaje transcrito). De hecho, por lo menos en las lenguas europeas, muchos signos se oralizan, es decir, se adaptan como glosas en el medio oral:

---

[31] Catach [30], "La escritura como plurisistema, o teoría de L prima" en Catach, ed., *op. cit.* [31] 1996, pp. 310.

[32] Rey-Debove, "En busca de la distinción oral-escrito" [117] en Catach, ed., *op. cit.* [31] 1996, p. 103; Nótese que este esquema no toma en cuenta fenómenos como el dictado que, según Rey-Debove, es un enunciado oral que debe facilitar la transcripción, es decir, *pretranscrito*.

> Estos signos no pueden ser asimilados a signos de puntuación, y no tienen equivalentes prosódicos; ":", que significa *por lo tanto, así, por ejemplo*; "()", que significa *por ejemplo, también llamado, que viene de, que se encuentra en..., que es... que vale...* y muchos otros: las bastardillas o las comillas también deben ser glosadas. una palabra X en bastardillas se lee [lapalabraekis], etc. Además, lo escrito contiene signos que no pertenecen al lenguaje (=, →, etc.) que deben traducirse claramente.[33].

Tampoco es extraño que, aun en lenguas de escritura alfabética, se incluyan símbolos no alfabéticos, desde los clásicos de una máquina de escribir (#, $, %, &, =, +, etc.) hasta los más variados *logogramas* (♡, ∝, ®, ©, ∀, ∃, ∈, ∉, etc.), para señalar los más diversos significados sin recurrir a la lengua propiamente. De hecho, como dice Haarmann, la escritura alfabética *moderna* no ha podido renunciar al componente adicional del principio logográfico[34].

Entonces, en el marco de la lengua escrita (cuando menos en el de aquella con escritura alfabética), podemos clasificar los signos de un corpus en tres tipos[35]: *alfagramas* (los símbolos alfabéticos[36] con o sin diacríticos), *topogramas* (los signos de puntuación, que incluyen a los espacios en blanco y las variantes de caracteres —versalitas, bastardillas, etc.[37]) y *logogramas*

---

[33] Rey-Debove, art. cit. [117] 1986, p. 107.

[34] Véase Haarmann, *Universalgeschichte der Schrift* [59], Campus, Francfort del Meno, pp. 207-210.

[35] Anis, art. cit. [9] 1986, p. 273.

[36] Como es bien sabido, los griegos inventaron el alfabeto al tomar los símbolos fenicios para representar por primera vez en la historia no solamente las consonantes, sino también las vocales; véase Mounin [107], *Historia de la lingüística desde los orígenes al siglo* XX, Gredos, Madrid, (*Manuales* 16) 1989 [1967], p. 91.

[37] Los signos de puntuación occidentales están tan presentes en nuestra cultura que son transparentes. Pero no nos podemos permitir asumir que funcionan igual en todas las culturas. De hecho, no sería extraño que su uso cambiara de una persona a otra. En parte por eso es importante estudiar seriamente este fenómeno, incluso en su dimensión histórica, ya que la presencia de topogramas en escrituras alfabéticas es casi tan antigua como el alfabeto mismo: si bien en un principio los griegos escribían sin interrumpir ni palabras ni oraciones, es en siglo IV a. de C. cuando, para señalar el inicio de un tema nuevo, comienzan a insertar una raya horizontal llamada παράγραφος (lat. *paragrăphus*). Luego, la tradición retórica empezó a dividir el discurso en segmentos de diferentes longitudes. Así, Aristófanes de Bizancio (inventor del sistema de acentuación del griego y de la colometría tradicional que pone claridad en las unidades métricas de la poesía (véase Reynolds y Wilson, *Copistas y filólogos* [118], Gredos, Madrid, 1986 [1974], pp. 16, 22), marcaba el final de una sección corta (fragmento), llamada κόμμα (lat. *cŏmma*), mediante un punto después de su última letra, al centro; el de una más larga (miembro), κῶλον (lat. *colon*), con un punto abajo; y el de la más larga, περίοδος

(unidades significativas, iconos, cifras, siglas, etc.).

Los signos de puntuación más difundidos pertenecen a la tradición europea de puntuación que, de hecho, ha sido transplantada a otras regiones del mundo junto con los diversos alfabetos descendientes del latín y del griego[38]. Pero no se trata de hacer un recuento detallado de los sistemas de puntuación del mundo[39], sino de establecer que, donde las lenguas se escriben —aun cuando su escritura no sea conocida por sus hablantes—, a menudo se recurre a señales de algún tipo cuando menos para marcar pausas, pero también para clarificar la estructura gramatical y, por lo tanto, el significado de los textos (es de notarse que las pausas y la estructura gramatical tienden a coincidir). Eso quiere decir que los córpora que puedan compilarse para cualquier lengua del mundo tendrán con toda probabilidad algún tipo de puntuación. Lo que es más, si se trata de córpora computarizados, será la europea, ya que sus signos de puntuación son parte de la cultura informática mundial.

(lat. *periŏdus*), mediante un punto arriba (Corominas y Pascual, *op. cit.* [44] 1991, *s.v.*, GRÁFICO, COMA I. CÓLICO, EPISODIO). Aunque sirvió de base al sistema moderno de puntuación, éste y otros sistemas se usaron en realidad poco. Luego, entre los siglos VII y VIII, cuando se distingue por primera vez entre mayúsculas y minúsculas, se empezaron a escribir las palabras de nuevo separadas, ahora mediante un espacio vacío, y a marcar el principio de las oraciones mediante una mayúscula. Además, algunas versiones modificadas de la puntuación de Aristófanes siguieron en uso. En comparación con estas convenciones, casi toda la puntuación durante la Edad Media se puede considerar errática tanto por su uso, como por su respeto a la sintaxis. En esencia, los signos de puntuación servían para indicar pausas de distintas longitudes. Finalmente, en el siglo XVII, Aldo Manuzio sugirió por primera vez en su *Orthographiae ratio* que el objetivo principal de la puntuación debería ser el esclarecimiento de la sintaxis. Así, para el siglo XVIII, ya estaban definidos para las lenguas europeas los principales signos de puntuación que hoy conocemos.

[38] En algunas lenguas, como el hebreo y el árabe, la puntuación europea fue adoptada con lentitud y en otras se utilizó desde la adopción de un sistema de escritura alfabética, como en muchas de las lenguas americanas. Otras lenguas, como el hindi, el bengalí y el chino, tienen otras marcas de fronteras entre oraciones o versos (una o varias rayas verticales, círculos vacíos, etc.) que también se usan junto con los signos de puntuación europea.

[39] En realidad los sistemas de escritura son innumerables, incluso donde los prejuicios culturales no permiten imaginarlo. Tómese, por ej., el caso del mito europeo del África "negra": frente a la imagen estereotipada de un África sin escritura, está la realidad de un inventario enorme de sistemas que se han utilizado desde la antigüedad hasta la fecha; véase Battestini, "Escrituras africanas (inventario y problemática)" [15], en Catach, ed., *op. cit.* [31] 1996, pp. 195-205.

Lo que importa recalcar es que todos estos topogramas, así como todos los otros símbolos o logogramas que puedan ocurrir en la cadena escrita, ya sea que sirvan como marcadores de pausas, desambiguadores de sintaxis o palabras plenas[40], tendrán muy poca probabilidad de ocurrir al interior de una unidad lingüística pequeña —tales como un afijo, o incluso una raíz[41]— y ocurrirán por lo regular en unos pocos lugares dentro de unidades mayores —en las fronteras entre sintagmas, oraciones o períodos.

También hay que recalcar que ningún corpus es lenguaje ni oral ni oralizado, sino escrito y transcrito. Además, como ya se vio, las transcripciones —como la escritura en general— presuponen algún sistema de puntuación que puede servir para, cuando menos, detectar fronteras lingüísticas de algún tipo. De todas maneras, para los córpora transcritos, se puede ignorar la puntuación en el cálculo de una glutinosidad "oral". Sin embargo, para los córpora de lengua escrita, parece natural auxiliarse del sistema de puntuación en el cálculo de una glutinosidad que deberá etiquetarse como "escrita".

Carácter asociativo de las dimensiones de la glutinosidad

En esta sección se resume la discusión de las dimensiones pertinentes a la glutinosidad y se examina la asociación entre las magnitudes que las miden. En concreto, se trata de justificar no solamente que las medidas de entropía y economía se pueden asociar para medir la di-

[40]También pueden fungir como palabras estructurales. Considérese por ejemplo '&', que en varias lenguas europeas se usa como equivalente de la conjunción 'y' (*and, et*, etc.).

[41]Nótese que es cuestión de probabilidad. Sin embargo, además de que de por sí no es tan raro que una palabra aparezca dividida en sílabas mediante guiones (lo que implica que tanto raíces como afijos pueden ser interrumpidos por dichos guiones), hoy en día no es extraño que ciertos topogramas e incluso logogramas aparezcan al interior de ciertos segmentos: por ej. 'c@lid@d' y 'StudentIn' (que en alemán corresponde simultáneamente a 'Student' y 'Studentin').

mensión de la glutinosidad, sino también que, para esto, son las mejores (independientemente de los procedimientos escogidos para su cálculo).

Es claro que las cantidades de entropía y economía son necesarias para definir un cálculo de la glutinosidad. Primero, porque obedecen a fenómenos verdaderamente universales del lenguaje. Segundo, porque son definitorios de los segmentos que se adhieren a otros para darle estructura al lenguaje (por lo que, como vimos en los capítulos anteriores, permiten descubrir unidades lingüísticamente pertinentes). Tercero, porque miden características de las relaciones de cada segmento con el resto de los segmentos del corpus (qué tanto se adhiere un segmento a los demás). Y cuarto, porque —aunque las dos medidas están, como se vio, muy relacionadas y, de hecho, co-varían junto con la glutinosidad— éstas no son lo mismo, es decir, miden cosas distintas.

Hay, por un lado, cierta interacción entre estas magnitudes: mayor economía puede implicar mayor entropía (a más opciones posibles, mayor incertidumbre de lo que sigue). De hecho, la medida de entropía de Shannon y Weaver crece con el número de opciones posibles. Pero mayor entropía no implica necesariamente mayor economía (pocas opciones posibles pueden causar una relativa gran incertidumbre). Aunque sean pocas las alternativas posibles, la entropía será alta cuando sean equiprobables (sin embargo, no sería extraña la situación en que muchos signos no tan equiprobables resulten en más entropía que pocos signos "muy" equiprobables).

Pero, por otro lado, los cuatro casos siguientes siempre son perfectamente posibles:

Alta economía, alta entropía Este es el caso de los segmentos más utilizados gramaticalmente. Se trata de un conjunto relativamente pequeño de segmentos muy frecuentes que ocurren antes (o después) de una gran variedad de signos posibles: la gran variedad que

acompaña a estos segmentos está compuesta de signos de probabilidades bajas y muy repartidas (son signos más o menos equiprobables) —los segmentos con valores altos de economía y entropía son los más gramaticales del sistema (entre los que se cuentan las partes invariables de la oración). En las tabla 3.7 al final del capítulo anterior, la tabla 4.8 al final de este capítulo y en la tabla D.1 del apéndice se listan las palabras gráficas del español mexicano con estas características. Destacan las conjunciones (por ej., 'y', 'o', 'que'), pronombres clíticos ('se', 'la', 'los', 'me', 'le', 'lo', etc.), artículos ('el', 'la', 'un', etc.), preposiciones (por ej., 'de', 'para', 'con'), etc., todos ellos con los valores más altos de *cliticidad* relativa en el corpus.

Alta economía, baja entropía En este caso, tendríamos un segmento después (o antes) del cual ocurren muchísimos signos, algunos de los cuales tendrían una probabilidad de ocurrir considerablemente mayor que los demás —los segmentos de alta economía y baja entropía son de carácter gramatical y aparecen pegados a otros de carácter también gramatical: son aquellos que marcan las fronteras de los sintagmas porque circulan en las órbitas más alejadas de los núcleos sintagmáticos. En español sería el caso de aquellos segmentos que preceden sintagmas nominales, por ejemplo, la preposición 'de' o la conjunción 'y' que típicamente aparecen antes de sintagmas que se inician con artículos u otros determinadores (en el corpus es mucho más probable que una frase nominal empiece con uno de estos que con un sustantivo). Así, en las tablas 3.3 y 3.4 del final del capítulo anterior los segmentos 'de' e 'y' exhiben los valores más altos de economía relativa, pero con valores de entropía bajos en comparación con aquellos de segmentos con índices de *cliticidad* menores o similares.

Baja economía, alta entropía Aquí se trata de un segmento después (o antes) del cual ocurrirían unos pocos signos equiprobables, es decir, cada uno con más o menos la misma probabilidad de ocurrir después (o antes) del segmento examinado —formas gramaticales no muy productivas por poco comunes, antiguas o restringidas a ciertos tipos de discursos. Por ejemplo, los morfemas de los paradigmas de conjugación correspondientes a la segunda persona del plural informal, tan productivo en España, pero tan restringido en América: 'os', *~áis*, *~éis*, etc. Son formas de uso tan restringido en el corpus que, como se apuntó en el primer capítulo, sus datos cuantitativos son incluso de carácter dudoso.

Baja economía, baja entropía El caso de glutinosidad nula o mínima sería aquel en que antes (o después) del segmento examinado hubiera poquísimos signos posibles, la mayoría de los cuales estarían caracterizados por precisamente lo contrario (alta economía y baja entropía); es decir, la ocurrencia del segmento en cuestión sería una garantía de que se supiera qué segmento ocurriría inmediatamente antes o después —son las formas de contenido alrededor de las cuales orbitan los segmentos gramaticales. Se trata de la mayoría de segmentos que ocurren en el corpus. Después de cada una de ellas es casi seguro que ocurra una forma gramatical (del tipo alta economía-alta entropía). Por ejemplo, en el sintagma **'y no ladró más'** alrededor del tema *ladr~* orbitan los segmentos *~ó*, 'no', 'más' e 'y', todos ellos con valores altos de economía y entropía.

Tabla 4.5: Selección de sufijos con más entropía que economía

sufijo	fr.	econ.	entrop.
~áis	3	0.2153	0.7263
~éis	4	0.67	0.63
~aos	1	0.2	0.76
~ei	6	0.3301	0.8325
~oides	4	0.4259	0.7539
~ed	3	0.4612	0.7882
~it	4	0.4807	0.7995
~arselas	3	0.4248	0.7224
~adisimo	3	0.4806	0.7157

De estos cuatro casos, solamente el último correspondería a los lugares donde no hay pegamento —al interior de los ladrillos sapireanos— o hay muy poco —entre sintagmas. El único medio formal que se ha explorado en esta tesis para distinguir entre los lugares donde hay poca glutinosidad, o no hay porque se trata del interior de un ladrillo sapireano, y aquellos donde hay poca, o no hay porque se trata de una frontera entre sintagmas, ha sido la puntuación.

En su carácter de efecto del discurso escrito, la puntuación no es precisamente un factor esencial al cálculo de la glutinosidad. Pero parece una herramienta muy útil para afinar la estimación de esta fuerza de adhesión. Además de no ser un factor esencial, la puntuación tiene algunas desventajas: por ejemplo, no son marcas de lengua oral, sino de lengua escrita o transcrita, y los topogramas varían de sistema en sistema y ni siquiera se puede decir que siempre coincidan con la sintaxis de las lenguas. Por otra parte, también tiene sus ventajas: por ejemplo, hace evidente el problema de la lengua escrita en el campo de los estudios de córpora y, por lo menos en el *CEMC*, es una marca más o menos confiable de frontera entre sintagmas.

Por todas estas razones, escogí —por un lado— las dimensiones de entropía I y economía

Tabla 4.6: Selección de sufijos con más economía que entropía

sufijo	fr.	econ.	entrop.
~s	12013	0.9968	0.4609
~r	2587	0.9482	0.4096
~do	2437	0.9473	0.4055
~mente	981	0.9758	0.3539
~ón	957	0.9197	0.3581
~rse	786	0.9031	0.2568
~mos	1103	0.857	0.4064
~te	1072	0.8838	0.4571
~ones	815	0.9561	0.4336
~sión	863	0.7837	0.4398
~da	441	0.9297	0.3343
~ron	317	0.9387	0.2056
~ndo	345	0.9018	0.2397
~rá	462	0.836	0.3499
~me	565	0.8561	0.4848
~res	334	0.9553	0.434
~nte	212	0.8639	0.2058
~dos	214	0.92	0.2998
~miento	248	0.7726	0.2388
~or	324	0.8663	0.4853
~l	358	0.8039	0.4743

S (de hecho, $\frac{S}{S}$, ya que se trata de una cantidad de signos de un nivel por cada signo del nivel anterior) como las dimensiones esenciales de la dimensión glutinosidad G y —por el otro— las cantidades de puntuación (también dimensión S) como una magnitud auxiliar que nos permita aprovechar las ventajas del texto escrito en la tarea de segmentar el discurso. Después de este indispensable análisis de las dimensiones involucradas en la derivación de la glutinosidad, podemos por fin adentrarnos en la cuestión de las unidades con qué medirla.

4.3.3 Las unidades de medición

Este apartado trata de la cuestión de unidades dentro de la teoría de la medición en general y explora los tipos de unidades pertinentes a una glutinometría formal basada en la transmisión de información entre segmentos y en la necesidad de economía en el lenguaje.

Una unidad es un intervalo básico que se manifiesta igual como parte abstracta de una escala conceptual que como parte concreta de una escala material. Además, como ya se estableció arriba, las magnitudes se expresan mediante una función P de la propiedad que un objeto x tiene en la escala s: $P(x, s) = ru$, donde r es un valor numérico y u la unidad escogida. Así, la medición depende inevitablemente de la selección de unidades. Entonces, la propiedad de que un objeto lingüístico sea afijo, clítico o palabra gramatical (o la propiedad de adherirse en cierta dirección a otros objetos del corpus), podrá medirse en una escala conceptual con un origen absoluto (en cero) y un orden métrico de unidades o grados de glutinosidad, mediante un glutinómetro que lleva implícita una escala "material" (es decir, una representación de la escala conceptual) donde se reflejan estas unidades y un cero u origen que representa la ausencia de dichas unidades de medición.

Unidades primarias y derivadas

Las unidades primarias (como por ej. el metro, el segundo, etc.) miden magnitudes de dimensiones primitivas o fundamentales (longitud L, tiempo T, etc.). Es obvio, entonces, que las unidades derivadas miden magnitudes de dimensiones complejas o secundarias (el volumen $L \times L$, la velocidad L/T, aceleración L/T^2, etc.). El investigador es libre de escoger las unidades de medición de cada dimensión fundamental, a partir de las cuales todos los demás valores se determinan. De hecho, se logra una simplicidad considerable al escoger arbitrariamente solamente las unidades de las dimensiones fundamentales para obtener una base de dimensiones independientes y luego permitir que las dependencias conocidas determinen todas las otras unidades. Tal sistema de unidades se considera coherente. Todas las de la base se dicen primarias y todos los demás derivados o secundarios.

Como se dijo arriba, las magnitudes simples o fundamentales no requieren de ningún análisis teórico previo para determinar unidades. Al materializar un concepto primario mediante la selección de una unidad, más que descubrir una noción natural o crear un concepto absoluto, se está proponiendo una convención[42]. Según se definan, las unidades pueden ser naturales o artificiales[43]. Por otra parte, las magnitudes derivadas, como también se estableció arriba, sí requieren de un marco teórico y de un sistema coherente de unidades primarias. De hecho, se obtienen multiplicando y/o dividiendo las magnitudes fundamentales sin introducir factores numéricos y tomando en cuenta el análisis previo de estas magnitudes[44]. De esta manera, a partir del análisis de las magnitudes y dimensiones pertinentes llevado a cabo arriba, se presentan a continuación las unidades de la glutinometría. La tabla 4.7 resume la definición de estas unidades.

Unidades glutinométricas

Como vimos en los capítulos anteriores, en el marco de una glutinometría para lenguas como el español, las magnitudes tanto de transmisión de información como de economía entre segmentos están en proporción directa al grado de afijalidad o cliticidad de un segmento. Y, si hemos de medir la glutinosidad en términos de por lo menos estas dos dimensiones, $G = I \times \frac{S}{S}$ $(GL = hk)$, —es decir, en términos de información transmitida al interior de una estructura definida por el ahorro o economía de los signos lingüísticos— podemos uti-

[42]Bunge, *op. cit.* [25] 1967, p. 225.

[43]Generalmente, las unidades se definen en términos de algún fenómeno natural observable fácilmente reproducible y muy invariante (tal como la longitud de onda de alguna fuente de luz). Otro método es el de las unidades artificiales que se preservan como objetos materiales únicos y se conocen como unidades estándar (por ej. el metro en París).

[44]Bunge, *op. cit.* [25] 1967, p. 228.

lizar la unidad tradicional de información, el *bit* (véase su definición en el primer capítulo, página 85), y alguna unidad que represente la economía de signos inherente a la estructura en cuestión. Pero ¿qué es exactamente lo que hemos medido aquí como economía entre segmentos? En esencia, el cociente de de Kock nos da el número de segmentos que acompañan al signo examinado, dividido entre el número de segmentos que alternan con él. Es decir, la glutinosidad corresponde a la cantidad de *bits* transmitida en una estructura restringida por una economía de signos (proporcional al promedio de signos adyacentes a los miembros de un paradigma). A la unidad de aprovechamiento económico, *acompañante/alternante*[45], se le puede llamar *Kock*, en honor a Josse de Kock. De tal manera que la unidad de glutinosidad podría ser un *bit*×*Kock* y podría llamarse *Varrón* en honor al antiguo gramático latino. Así:

$$Varrón = bit \times Kock = \frac{bit \times \text{acompañante}}{\text{alternante}}$$

representaría la unidad de información transmitida entre un signo y sus posibles seguidores (o predecesores) para crear un signo del nivel siguiente por cada signo que alterne con el primero. Es decir, la cantidad de *bits* por unidad de "aprovechamiento" de signos en una estructura que funciona como signo de otro nivel.

Esto implica una escala de números reales $\mathbb{R}$ entre 0 y el infinito, $[0, \infty]$, ya que ninguna de las dimensiones puede ser negativa y los signos nuevos son una clase abierta. De esta manera, la glutinosidad se medirá probablemente en *kilovarrones* (*kV*) y es de esperarse que la afijalidad tenga valores más altos que los de la cliticidad.

---

[45] También *(signo nuevo del nivel superior)/(alternante del nivel inferior)*, porque en una relación económica, cada combinación del signo examinado con cada segmento acompañante constituye un signo nuevo del nivel siguiente, similarmente cada signo que alterna con el examinado constituye con sus segmentos acompañantes nuevos signos del nivel superior.

Unidades glutinométricas y lengua escrita

Como quedó establecido arriba, una de las diferencias entre lengua hablada y lengua escrita es la puntuación, la cual puede o no proporcionar alguna estrategia de demarcación visual (mediante espacios) de las palabras según la cultura a la que pertenece dicha lengua. De hecho, al adoptar la premisa de que una palabra sea aquello que aparece entre dos espacios (véanse las premisas de este trabajo en la página 14 de la introducción), se están dejando intervenir las convenciones de la lengua escrita en el fenómeno más general del habla. Pero esto es en gran medida inevitable al trabajar con córpora, incluso si se trabaja solamente con transcripciones de lengua hablada, porque quien transcribe en realidad traduce las estructuras que oye al sistema de la escritura. Así, una transcripción puede ser tan elaborada como una que incluya ruidos, pausas e interpretaciones sintácticas de quien transcribe o tan sencilla como una secuencia de palabras delimitadas cuando menos mediante espacios (cosa que también implica la interpretación de quien transcribe)[46].

Por eso, para llevar a cabo una glutinometría del habla, no hay que olvidar el problema de los límites de la palabra y, por lo menos, dejar claro que las medidas resultantes de un procedimiento glutinométrico deben tomarse con cautela si el objeto es la lengua transcrita y se utilizan criterios culturales para delimitar las palabras o, en su defecto, establecer claramente los criterios que se utilicen.

Sin embargo, la glutinometría bien puede aprovechar las características de la lengua escrita (más allá de tomar en cuenta los espacios como fronteras de palabras) para determinar las

[46] Cabe notar que la ausencia de espacios en una cadena de caracteres no tiene por qué ser un obstáculo en la segmentación de esa cadena en unidades lingüísticas: se puede determinar una ventana de varios símbolos a un lado y a otro y medir los índices pertinentes a partir de dichos símbolos.

fuerzas de atracción entre los objetos lingüísticos representados gráficamente. Por ejemplo, el número de signos de puntuación asociado a los vocablos. Pero en ese caso tendría que hablarse explícitamente de glutinometría de lengua escrita y no simplemente de glutinometría. De esta manera, para conocer qué tanta glutinosidad por signo de puntuación hay a cada extremo de un segmento, podemos utilizar la unidad siguiente:

$$\frac{\textit{Varrón}}{\text{topograma}}$$

Es decir, a más topogramas, menos glutinosidad y, por lo tanto, más probabilidad de frontera sintagmática. Y, a menos topogramas, más cliticidad. Recuérdense, por ejemplo, las tablas 3.3 y 3.4 presentadas al final del capítulo anterior. Como ya se explicó (en la página 200), la tercera columna alberga el complemento de la probabilidad que cada segmento tiene de ocurrir con un signo de puntuación ($1 - \frac{r_x}{f_x}$) (índice de no puntuación). Así, en la tabla 3.3, vemos que el 99.77% de las 33,623 ocurrencias del segmento 'se' (núm. 1) no fueron seguidas por ningún topograma, es decir, sólo 78 (0.33%) ocurrieron inmediatamente antes de algún signo de puntuación. Cada uno de esos signos es evidencia de glutinosidad disminuida, pero el altísimo porcentaje de ocurrencias sin topogramas asegura una cantidad de fuerza glutinosa muy respetable. En cambio, en la misma tabla tenemos al segmento 'bien' (núm. 75) con un porcentaje de 26.51% (100%-73.49%) de ocurrencias seguidas por un topograma, indicio importante de que después de 'bien' puede haber frontera de sintagma, por lo que su fuerza de adhesión a los segmentos siguientes debe verse disminuida.

Nótese que el espacio en blanco (' ') cuenta como topograma, por lo que al exterior de la palabra, nunca se estaría dividiendo la glutinosidad entre menos que uno. Por otra parte, esta unidad no sería aplicable al interior de la palabra gráfica, ya que casi siempre estaríamos

dividiendo entre cero. Además, es obvio que de haber algún topograma (guión, diagonal, etc.) al interior de una palabra, su presencia contaría como evidencia de segmentación (por lo menos en español y otras lenguas como el portugués, donde se acostumbra adherir ciertos segmentos a ciertas bases con la presencia de guiones; por ej. 'socio-económico'[47]) por lo que el número de topogramas estaría en relación directamente proporcional (multiplicativa) con la glutinosidad. Todo esto sólo hace evidente que por lo menos al interior de la palabra gráfica, todavía hace falta estudiar la distribución de topogramas. Por lo tanto, esta propuesta de glutinosidad "escrita" está restringida entonces al exterior de la palabra gráfica.

En la tabla 4.7 se resume la definición de las unidades glutinométricas. Las dimensiones involucradas son, como se dijo arriba, la información o entropía I (comúnmente medida en *bits*) y la cantidad de signos económicos $\frac{S}{S}$ (ó $\frac{S}{S^2}$). Ésta última dimensión corresponde a varias magnitudes: cantidad de signos nuevos del nivel superior (relación sintagmática), cantidad de signos del nivel inferior involucrados en la formación de los primeros (relación paradigmática) y cantidad de topogramas entre dos signos de un sintagma. La magnitud de economía (número de signos nuevos por cada signo del nivel anterior) correspondería a la dimensión $\frac{S}{S}$ (signo económico). Al tomar en cuenta los topogramas en la medición de economía de signos, la dimensión correspondiente sería $\frac{S}{S^2}$ (signo económico de lengua escrita). Por último, la dimensión de glutinosidad correspondería a la dimensión de economía multiplicada por la de información esperada en la transición de un signo a otro dentro del sintagma: $\frac{S \times I}{S}$ y $\frac{S \times I}{S^2}$.

[47] Esto implica, por supuesto, la presunción de que son relativamente pocas las ocurrencias de guiones para separar sílabas, cosa que *a priori* no parece muy cierta.

Tabla 4.7: Unidades de la Glutinometría

unidad	tipo	dimensión	descripción
signo nuevo (*acompañante*)	primaria	S (signos en relación sintagmática)	signo del nivel superior formado mediante el signo cuya glutinosidad se mide y los segmentos que le acompañan
alternante	primaria	S (signos en relación paradigmática)	signo en distribución complementaria con el signo cuya glutinosidad se mide
topograma	primaria	S (puntuación)	signo de puntuación que ocurre como acompañante del signo cuya glutinosidad se mide
Kock	derivada	$\frac{S}{S}$ (economía)	signo económico: mientras más acompañantes y menos alternantes, mayor economía: $Kock = \frac{signo\ nuevo}{alternante}$
bit	primaria	I (entropía)	unidad de sorpresa que causa el no saber qué signo acompaña al signo cuya glutinosidad se mide
Varrón	derivada	$G = \frac{S \times I}{S}$ (glutinosidad)	unidad de sorpresa o información por signo económico: $Varrón = bit \times Kock = bit \times \frac{signo\ nuevo}{alternante}$
$\frac{Varrón}{topograma}$	derivada	$G = \frac{S \times I}{S^2}$ (glutinosidad "escrita", sólo entre palabras gráficas)	unidad de sorpresa por signo económico de lengua escrita: $bit \times \frac{signo\ nuevo}{alternante \times topograma} = \frac{bit \times Kock}{topograma}$

4.3.4 Las bases axiomáticas de la medición

En esta sección se hacen explícitos los axiomas, reglas o premisas básicas, que constituyen los cimientos de una teoría de la medición de las relaciones de asociación entre los objetos lingüísticos de un corpus de lengua natural. Con esto en mente, se examinan los tipos de axiomas característicos de los procesos de medición, para luego formular aquellos pertinentes a una glutinometría.

Específicamente, los axiomas son principios indemostrables pero evidentes[48] y constituyen sistemas que varían mucho en tamaño y estructura, según el fenómeno que se desee observar. Es decir, los sistemas axiomáticos difieren entre sí por la cantidad y estructura que los caracteriza. Además del principio de no contradicción, los axiomas de medición incluyen cuando menos una relación de orden. Pero, para la mayoría de los fenómenos, los axiomas de orden no son suficientes para formular muchas leyes científicas, porque no requieren una representación numérica determinada con suficiente detalle. Por eso, para matizar al fenómeno observado, es decir, darle mayor estructura y singularidad a su representación, se incluyen otros axiomas que permiten realizar operaciones empíricas, tales como la adición de propiedades de entidades físicas (que tienen varios componentes independientes) o de otros primitivos que conducen a una representación geométrica. En suma, los tipos de axiomas posibles son de orden, extensión, de asociación o unión y geometría, que se analizan a continuación.

Los axiomas de orden están relacionados con el concepto de escala que, como quedó establecido al principio de esta sección, es un complejo que comprende tanto los hechos como las ideas de un fenómeno: hay una escala conceptual para cada escala material y las dos presuponen sendas relaciones de ordenamiento ($\dot{\geq}$ es la relación concreta de ordenamiento entre los objetos de un fenómeno y $\geq$ es la relación conceptual de la escala numérica). Estos axiomas garantizan que el orden impuesto a los objetos al asignarles números sea el mismo orden de los datos observados o medidos. Es decir, dado un proceso de medición, el orden inducido en los objetos por las medidas asignadas es el mismo orden observable empíricamente en dichos objetos. Esto significa que toda desigualdad numérica que surja en el proceso de

[48] Abbagnano, *op. cit.* [1] 1991, *s.v.*

medición es transitiva ($x \geq y$ y $y \geq z$ implican que $x \geq z$) y representa una relación en un solo sentido ($x \geq y$ ó $y \geq x$, pero no ambas). De la misma manera, se espera que la desigualdad empírica sea transitiva y en un sentido. Para distinguirla de la numérica, ésta se representa mediante el símbolo '$\dot{\geq}$' (con un punto arriba). Esta relación no implica que el que dos objetos tengan el mismo valor, sean uno mismo. Este tipo de relaciones se conocen generalmente como de orden débil, lo que significa simplemente que aunque x y y sean iguales, como se dijo arriba, no son necesariamente lo mismo. Otras propiedades numéricas que el orden empírico debe reflejar es, primero, que entre dos números racionales distintos siempre exista otro número racional y, segundo, que a todos y a cada uno se les pueda asignar un número entero (que sean contables)[49].

Los axiomas de extensión están relacionados con magnitudes y dimensiones *aditivas*. Ejemplos clásicos son los atributos de duración, longitud y masa que, al estar presentes en varios objetos, se pueden combinar o concatenar para formar nuevos objetos o eventos que también exhiben el atributo en cuestión. En general, las cantidades se pueden clasificar en por lo menos dos grupos, según se trate de magnitudes aditivas o no —lo que significa que la experiencia demuestra que las propiedades de los objetos x y y se pueden sumar, es decir, existe para esta propiedad una operación física de adición '$\dot{+}$' (que en el plano simbólico corresponde simplemente a '+'). De acuerdo con esto, las cantidades se llaman extensivas cuando son aditivas, como la distancia (la longitud de dos palos yuxtapuestos es la longitud de los palos combinados) y la duración (tanto simbólica como empíricamente, dos minutos

[49]No hay axiomas universales para el proceso de contar (Bunge, *op. cit.* [25] 1967, p. 215), pero sí hay muchas maneras de hacerlo. El problema principal es decidir cómo. Y siempre existe la posibilidad de comparar técnicas, (especialmente en cuentas indirectas) para discernir cuáles son mejores o cómo mejorarlos.

más tres minutos es igual a cinco minutos[50]), o intensivas, cuando en el nivel físico no pueden someterse a operaciones matemáticas (como las propiedades estadísticas y la temperatura[51]).

Los axiomas asociativos o de unión establecen que aquellos atributos que no se pueden medir empíricamente (es decir, directamente en el objeto) se pueden medir al observar la manera en que otras dimensiones, que puedan considerarse componentes de dichos atributos, cambian una en relación con las otras. Este es el tipo de axioma necesario para la medición de atributos tales como el hambre, la utilidad, la inteligencia, la velocidad, la afijalidad, etc. en términos de otros atributos que sí se pueden medir empíricamente. En pocas palabras, este tipo de axioma es el que nos permite concebir la glutinosidad (y, por lo tanto, la afijalidad, la cliticidad y, de alguna manera, la "gramaticalidad") en términos de cuando menos entropía y economía. La idea es que hay conceptos como éste que, para poderlos cuantificar y medir, es necesario encontrar correlaciones con otras propiedades cuantitativas, en lugar de tomarlo como una noción básica e irreducible[52].

Por último, los axiomas de geometría se ocupan de la representación de atributos dimensionalmente complejos mediante grupos de dos o más números. Se trata de la representación de puntos mediante coordenadas cartesianas en un plano (pares de números) o en un espacio

---

[50]Hay magnitudes que no son del todo aditivas, sino solamente bajo ciertas condiciones: por ejemplo, la probabilidad no es aditiva en cualquier par de eventos (mientras que la probabilidad de que ocurra en el discurso una vocal cerrada sí se puede sumar a la probabilidad de que ocurra una abierta —en el cálculo de la probabilidad de las vocales—, de ninguna manera se suma la probabilidad de que llueva hoy con la de que vaya a haber un accidente automovilístico mañana); tampoco cantidades físicas de volumen, masa y energía son del todo aditivas (el volumen de un litro de agua mezclado con un litro de alcohol no es igual a la suma de los volúmenes de cada litro por separado), por lo que a estas últimas se les llama *cuasiextensivas* (*ibid.* [25], p. 200).

[51]La temperatura del agua de una alberca no es la suma de la temperatura de cada litro de agua), lo cual no quiere decir que no se puedan llevar a cabo en el nivel simbólico ($5°C + 2°C = 7°C$ no es una falsedad), porque todas las cantidades sean o no extensivas son conceptos cuantitativos.

[52]*Ibid.* [25], p. 204.

(grupos de tres números, uno para cada dimensión en el espacio). La importancia de este tipo de axiomas está en que son cruciales en el diseño de instrumentos de medición, porque vinculan al mesurando con las magnitudes geométricas pertinentes[53]. Es obvio que este tipo de axiomas son pertinentes a una glutinometría para representar gráficamente las relaciones de las dimensiones involucradas en el cálculo de la glutinosidad y, por ende, las unidades con que se mida.

Como se puede ver, hay varias cuestiones "evidentes" que merecen hacerse explícitas en la concepción de un esquema de medición. A continuación, se examinan estas cuestiones en el marco glutinométrico, se trata de establecer que los objetos de la glutinometría son empíricamente distintos y contables, que sus relaciones se pueden cuantificar (por lo que los objetos se pueden ordenar según sus características formales) y que la glutinosidad es, en efecto, una magnitud derivada a partir de las cantidades de entropía y economía entre los objetos.

Un corpus es una construcción de carácter cultural en el que ocurren linealmente diversos objetos que acarrean ciertas cantidades de información y que se definen no por sí mismos sino por sus relaciones económicas con el resto de los objetos que ocurren en el corpus. Esos objetos se pueden delimitar mediante, cuando menos, alguno de los siguientes factores:

Cultural: las tradiciones escritas recurren a diversas estrategias de segmentación del discurso (por ej., la puntuación).

Psicológico: los hablantes de una lengua son capaces de segmentar el discurso siguiendo su intuición, su idiosincrasia o, inevitablemente, su interpretación de la tradición.

Estadístico: hay diversas medidas estadísticas que permiten segmentar el discurso (aquellas de no asociación entre digramas —prueba de χ^2, razón de semejanza, etc.—, pero también las de abajo).

---

[53] *Ibid.* [25], p. 236.

De frecuencia: los segmentos más frecuentes delimitan o estructuran al discurso.

De cuadros, hexágonos, etc.: los segmentos que ocurren en más estructuras combinatorias también estructuran al discurso.

Económico: ciertos segmentos ocurren profusamente para multiplicar las estructuras de información posibles y, por lo tanto, contribuyen al ahorro en cuanto al número de segmentos que los hablantes tienen que memorizar, y

Entrópico: los segmentos que ocasionan incertidumbre en cuanto a lo que sigue —aquellos seguidos por otros de mayor equiprobabilidad— estructuran al discurso.

Es claro que cada una de estas estrategias puede resultar en diferentes segmentaciones de un mismo corpus, pero lo importante es que todos estos factores son susceptibles de cuantificarse y, por lo tanto, sirven para descubrir diversas versiones de los objetos del corpus, ya sean de carácter especulativo (cultural y psicológico) o formal (frecuencia, economía, entropía, etc.). Es decir, que un corpus contiene miembros empíricamente distintos, que se distinguen entre sí y, por lo tanto, están separados y se pueden contar. Además, los objetos así delimitados no necesariamente corresponden —aunque se espera que se aproximen— a los ladrillos sapireanos, que son los objetos ideales a los que convergen las estrategias formales.

Axioma 1. Para cada muestra Ψ de objetos lingüístico-culturales, $o_1, o_2, o_3, \ldots, o_\xi$, existe una secuencia $x_1, x_2, x_3, \ldots, x_{\xi'}$ (o varias) de objetos definidos formalmente (es decir, cuantitativamente) que constituyen aproximaciones a la secuencia de objetos lingüísticos verdaderos $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{\xi''}$ que conforman Ψ.

De esta manera, cada uno de estos objetos mantiene diversos tipos de relaciones con el resto de los objetos del corpus y estas relaciones, como se dijo arriba, son cuantificables. Esta cuantificación implica la posibilidad de ordenar estos objetos según los valores que se obtengan para cada estrategia de delimitación. Por un lado, las estrategias especulativas (que,

si recurren a algún tipo de cuantificación, se trata de una contabilidad de simples impresiones) apenas permiten distinguir entre unas cuantas sombras de intensidad de asociación entre elementos que van desde aquellos con más intensidad (como afijos y clíticos), pasando por los menos asociados, pero estructuralmente significativos (pre- y posposiciones, etc.), hasta los más independientes (palabras plenas). Por otro lado, las estrategias formales o cuantitativas ofrecen tantos grados de asociación que un ordenamiento más sofisticado se vuelve posible. En concreto:

Axioma 2. Los objetos lingüísticos del corpus constituyen un sistema de relaciones $\dot{\mathbf{G}} = \langle \dot{G}, \dot{\leqq} \rangle$, donde $\dot{G} = \{\dot{g}\}$ es el conjunto de grados de glutinosidad y $\dot{\leqq}$ es la relación concreta de ordenamiento según estos grados. Este sistema concreto corresponde directamente al sistema conceptual $\mathbb{R} = \langle R, \leqq \rangle$, compuesto del subconjunto R de números reales y la relación aritmética de ordenamiento $\leqq$. Es decir, $\dot{\mathbf{G}}$ y $\mathbb{R}$ son sistemas isomórficos.

Como lo demuestran las diferentes estrategias para descubrir objetos lingüísticos, las relaciones entre éstos son de varios tipos. Algunas de ellas capturan la esencia de lo lingüístico mejor que otras. Por ejemplo, la propiedad que cada objeto tiene de adherirse a los demás objetos, ya sea en calidad de afijo, clítico o palabra gramatical —propiedad hasta ahora de tipo cualitativo (concebida desde la antigüedad en Asia y el Mediterráneo)—, es directamente proporcional a las cantidades de sorpresa que los acompañan (no que causen) y al número de signos nuevos que contribuyan a crear en el siguiente nivel, e inversamente proporcional al número de signos con que aparecen en distribución complementaria (con que alternan) en la creación de otros nuevos signos del nivel siguiente:

Axioma 3. A cada lado de cada objeto lingüístico de un corpus Ψ hay una fuerza de adhesión o glutinosidad (direccional, porque es particular para cada lado):

$$GL(x_i) = \frac{h_i \star \textbf{nuevos}_i}{\textbf{alternantes}_i} = h_i k_i.$$

donde x_i, h_i y k_i representan respectivamente al objeto formal, la sorpresa que le acompaña y el ahorro que le significa al sistema (al formar parte de un conjunto de *alternantes* que contribuyen a la formación de *nuevos* signos del nivel siguiente).

La relación entre economía y entropía como dimensiones a partir de las cuales se deriva la glutinosidad se puede representar gráficamente mediante puntos en un plano, como en la figura 4.4. Esta figura representa la cantidad de glutinosidad mediante puntos en un plano cartesiano. Nótese que cada dos puntos definen un área, que corresponde a la glutinosidad del objeto examinado (al que se le midió la entropía y la economía). Nótese también que cada bit por signo producido económicamente se representa en el plano cartesiano como un área de 1×1, que constituye la unidad de medición idónea para dicha energía de adhesión lingüística (que llamamos Varrón en el subapartado anterior).

Axioma 4. Las cantidades de glutinosidad corresponden a puntos en el espacio cartesiano, donde los ejes representan las dimensiones de economía (en signos producidos por cada signo alternante) y entropía (en bits). De esta manera, el área definida por cada par de valores de estas dos dimensiones representa la cantidad de glutinosidad asociada al extremo en cuestión del objeto examinado.

La glutinosidad no es aditiva en todos los contextos. No está muy claro que la fuerza de adhesión de un objeto se pueda sumar a la fuerza de adhesión de otro objeto para obtener

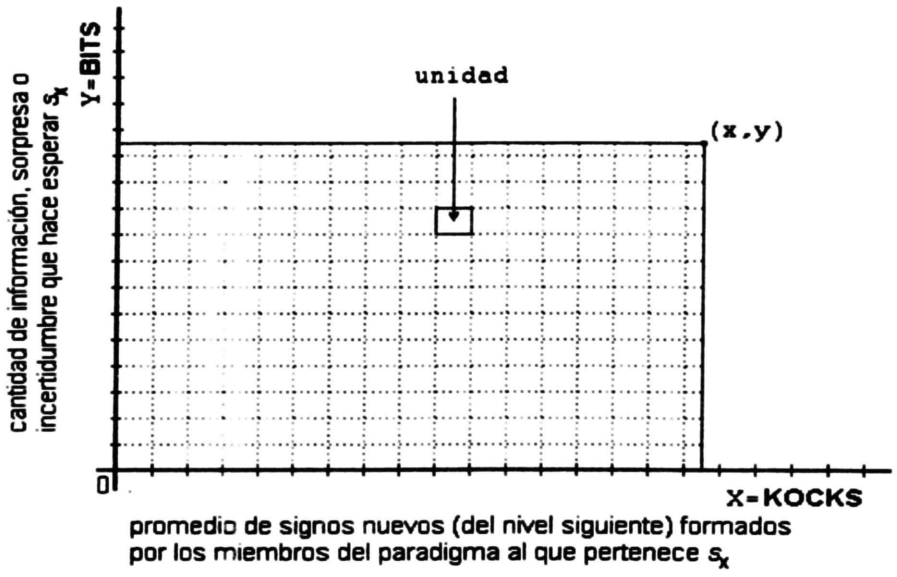

Figura 4.4: La glutinosidad en el plano cartesiano

la fuerza combinada de adhesión de los dos objetos. Lo que sí queda claro es que la fuerza de adhesión de, por ejemplo. los pronombres clíticos a verbos plenos no es la suma de la glutinosidad de cada clítico. Si acaso, se podría pensar en un promedio como el valor representativo de todos ellos. En ese sentido la glutinosidad no es extensiva. Sin embargo, la combinación de las dos cantidades de un mismo objeto sí parece resultar en algo más o menos coherente: al acumular la glutinosidad de un lado con la del otro para caracterizar a un objeto examinado (los segmentos más usados gramaticalmente deben tener los valores más altos), o, sobre todo, al compararlas (es decir, restarlas) para determinar en que dirección y qué tanto se adhiere un clítico.

Axioma 5. La glutinosidad es condicionalmente extensiva. Es decir, existe una operación empírica de adición '$\dot{+}$', que en el plano simbólico corresponde a '+', válida sólo para los valores de glutinosidad de un mismo objeto lingüístico: si

x es el extremo derecho de un objeto y y es el izquierdo del mismo, entonces $G(x\dot{+}y) = G(x) + G(y)$ (y, por lo tanto, $G(x\dot{-}y) = G(x) - G(y)$).

La utilidad más obvia de este axioma está sobre todo en la fundamentación para determinar los clíticos verdaderos (es decir, para distinguirlos de las palabras más gramaticales), pero todavía deben examinarse las consecuencias de la posibilidad de sumar los dos valores de un objeto: por ejemplo, no los clíticos, sino los nexos más gramaticales obtendrán los valores más altos[54].

Antes de continuar con la última parte de este apartado, es necesario resaltar que la importancia de enunciar explícitamente estas verdades indemostrables pero evidentes reside en que se hace patente que, pongámosle el nombre y los adjetivos que sean, cada objeto lingüístico de un corpus tiene algo (una "energía") a cada extremo que lo define y que se puede medir de diversas maneras, entre las cuales, los cálculos de entropía y economía son indispensables.

4.3.5 El problema del error

En este apartado se examinan los tipos de error inherentes a los procesos de medición. Una de las ventajas de cualquier estrategia de medición, como la glutinometría, es que, al igual que en la estadística y otros aparatos formales para conocer el mundo, el problema del error se tiene que tomar en serio. En lingüística, el término "error" se ha aplicado por ejemplo en el estudio de adquisición de segundas lenguas para designar las equivocaciones de

[54]Un clítico está muy asociado a un lado y no al otro, así que la suma de glutinosidad de los lados no puede ser tan alta como la de un nexo con glutinosidades altas para ambos lados.

los hablantes (*error analysis*[55]), donde, a grandes rasgos, se trata del estudio de las maneras y las causas en que los hablantes se equivocan. También se habla de "error" sobre todo cuando se aplican herramientas estadísticas que, como ya se dijo, implican un estudio sistemático de los errores. Fuera de eso, en los estudios del lenguaje, el uso del concepto de error en realidad no ha sido objeto de interés; es más bien un término de descalificación que un motivo de investigación[56].

Con respecto a los procesos de medición, en alguna época se creyó que los errores se podrían eliminar mediante el refinamiento tanto de principios científicos como de instrumentos para llevarla acabo. Pero actualmente se ha abandonado esta creencia y casi todos los reportes de mediciones de las más diversas índoles se acompañan de alguna indicación del grado probable de error o de las limitaciones de exactitud o precisión de dichas medidas. De hecho, todo proceso de medición está cargado de errores, por lo que el problema del error es uno de los temas centrales de la teoría de la medición. Además, toda "buena técnica de

---

[55]Bußmann, *op. cit.* [28] 1990, *s.v.* (FEHLERANALYSE, FEHLERLINGUISTIK).

[56]Como comentario histórico, eso no quita que el problema del error sí se haya contemplado desde muy temprano en la historia de los estudios del lenguaje. Por ejemplo, Varrón reconoció este problema (lat. *mendo*) en la exploración del origen de las palabras:

> [la etimología y la semántica] Ambos aspectos presentan bastante oscuridad, ya que no se conserva todo el proceso de acuñación de las palabras: el paso del tiempo ha hecho caer a algunas en el olvido. Además, ni todas las que se conservan están exentas de error [en el original *sine mendo*; en la versión de Kent, (*op. cit.* 1938, p. 5), "without inaccuracy"], ni tampoco las acuñadas correctamente se mantienen siempre íntegras, pues muchas palabras aparecen modificadas por alteraciones de letras; ni todo el origen de nuestra lengua deriva de palabras vernáculas. [..] Dado que en el lenguaje corriente estas palabras modernas y antiguas se ven sometidas a todo tipo de alteraciones, aquél que sea capaz de establecer de cuántas maneras puede producirse el cambio verá altamente facilitada su labor de exploración del origen de las palabras. Llegará a la conclusión de que los cambios experimentados se han debido —como he explicado en los libros precedentes— especialmente a dos grupos de cuatro causas: a) 1°, a la supresión o, 2°, adición de palabras, 3°, y a causas de su transposición o, 4°, de su transformación; b) pero también, 1°, al alargamiento o, 2°, a la abreviación; 3°, a la adición o, 4°, a la supresión de sílabas [Varrón *op. cit.* [133] 1990 [*ca.* 40 a.C.], V, 3-6, pp. 5-7].

Nótese que al tratar de explicar de qué manera las palabras se ven alteradas con el paso del tiempo y tratar de enunciar las causas de dichos cambios, los errores de los que habla se convierten más en parte de un fenómeno (ya no son errores propios) que en obstáculos infranqueables, por lo que no son motivo para abandonar su investigación de las palabras antiguas: "Por mi parte, si no pudiera seguirles la pista no andaré perdiendo el tiempo; en cambio, si pudiese, me apresuraré a seguirla" (*ibid.* [133], V, 5, p. 7).

medición demanda *escepticismo continuo* acerca de la exactitud de los resultados"[57].

Solamente en lo que a la medición se refiere, hay muchos tipos de errores, entre los que destacan los inherentes a los procesos de observación y muestreo, además de aquellos de tipo teórico. Los más importantes de este último tipo son los errores del marco teórico, es decir aquellos causados por asumir premisas falsas. En cuanto a los procesos de observación, se pueden clasificar por lo menos cuatro tipos de errores[58]:

1. Instrumentales: errores causados por la inexactitud de los instrumentos debida, por ejemplo, al desgaste.
2. Personales: errores debidos a la actuación de los observadores.
3. Sistemáticos: errores que se comportan de manera predecible, por lo que sus causas son deducibles.
4. Aleatorios: errores imprevistos cuyas causas no se conocen.

Aparte de los errores de observación, los errores más importantes son quizá los originados en el muestreo. Este tipo de error es inevitable, pero como regla general se asume que mientras más grande sea la muestra menor será este tipo de error[59].

Otra manera de clasificar errores se refiere a su inmediatez, es decir, al uso directo o indirecto de medidas erróneas en el cálculo de otras medidas. De esta manera, son errores directos aquellos involucrados en el proceso mismo de medición, mientras que son indirectos, aquellos en los que se *basa* el proceso de medición; por ejemplo, errores en el cálculo de la

---

[57]Cooper, *Instrumentación electrónica y mediciones* [43], Prentice Hall, Bogotá, 1982, p. 3.

[58]La clasificación es, por supuesto, relativa. Así, mientras Cooper (*op. cit.* [43] 1982, pp. 6-9) reconoce tres (brutos, sistemáticos y al azar), Bunge (*op. cit.* [27] 1998, pp. 272-274) señala dos (sistemáticos y aleatorios) que comprenden los cuatro tipos.

[59]Véase la discusión de muestras representativas en la lingüística estadística en Manning y Schütze, *op. cit.* [93] 1999: "having more training text is normally more useful than any concerns of balance, and one should simply use all the text that is available", pp. 119-120.

velocidad de la luz o de la masa de la tierra en los que se base algún proceso de medición causarán un error indirecto en dicho proceso.

Pero aparte de las maneras de clasificar los errores, existe toda una teoría de los errores que ha recibido mucha atención en el desarrollo de la ciencia y de las matemáticas, en especial de la teoría de la probabilidad. Aunque es debatible si todos los tipos de errores siguen siempre la distribución normal (considérese que un error sistemático bien puede tener alguna otra distribución —de hecho, sería sistemático debido a una distorsión regular que evitaría una convergencia normal), los errores al azar o aleatorios se distribuyen, por definición, normalmente[60].

Estos últimos son parte del fenómeno (es decir, no son propiamente errores) y siempre van a interferir (en el mejor de los casos, los verdaderos errores —factores que distorsionan los hechos debido a irregularidades mecánicas o humanas— son corregibles) con todo proceso de medición. En términos de Bunge, casi siempre[61] hay una discrepancia entre los valores reales $\dot{R}$, por un lado, y los valores medidos $\{m(\dot{r})\}$, por el otro. Y por lo general esta discrepancia es desconocida y rara vez se puede determinar. Esto se suele representar de la siguiente manera[62]:

$$m(\dot{r}) = \frac{p}{q} \pm \sigma$$

donde p y q son números enteros (los valores medidos son números racionales, es decir, todo

[60] Para exposiciones más detalladas de la teoría de la probabilidad de los errores que se comenta brevemente en los siguientes párrafos, véanse Bunge, *op. cit.* [27] 1998, pp. 273-274, Cooper *op. cit.* [43] 1982, pp. 9-17 y Woods *et al.*, *op. cit.* [136] 1986, pp. 80-102. En este último, se desarrolla con detenimiento el procedimiento para determinar intervalos o límites de confidencia.

[61] Solamente las operaciones más sencillas de conteo están libres de error.

[62] Bunge, *op. cit.* [25] 1967, p. 209.

proceso de medición resulta en aproximaciones fraccionarias) y $\sigma \geq 0$ es la diferencia entre el valor medido y el valor real.

Pero, aunque esta discrepancia sea casi siempre desconocida y —de hecho— no se pueda determinar, no implica que no se puedan estimar límites fuera de los cuales el error sería poco probable. Entonces, si la discrepancia $\Delta = r - m(\dot{r})$ es casi siempre desconocida, se puede determinar el error de un acto aislado de medición al calcular la diferencia entre el resultado de ese proceso aislado y un promedio de mediciones previas (mientras más mediciones previas, mejor promedio):

$$\varepsilon_i = m_i(\dot{r}) - \bar{m}(\dot{r})$$

donde $\bar{m}(\dot{r})$ es el promedio de valores medidos previos (u otro tipo de valor teórico):

$$\bar{m}(\dot{r}) = \frac{m_1(\dot{r}) + m_2(\dot{r}) + m_3(\dot{r}) + +m_N(\dot{r})}{N}$$

La esperanza es que el promedio $\bar{\varepsilon}$ de las discrepancias entre $m_i(\dot{r})$ y $\bar{m}(\dot{r})$ se aproxime a cero según crezca el número de observaciones. Esto sustentaría la suposición de que $\bar{m}(\dot{r})$ se acerca al valor $\dot{r}$ del mesurando —es decir, que converge hacia el valor *real*. Sin embargo, las únicas razones que nos permiten presumir esto son mucha fe y mucha buena voluntad. Pero es precisamente esta esperanza la motivación para encontrar y hacer operativas y comparar entre sí nuevas y mejores técnicas de medición[63]. El caso es que es improbable que este promedio, por representativo que sea del mesurando, ocurra en un acto aislado de medición (en parte porque nunca se mide la misma cosa dos veces con el mismo aparato[64]). Además, no se trata de un sólo número, sino de dos: el promedio $\bar{m}$ y su σ (error estándar).

[63] *Ibid.* [25], pp. 209-210.

[64] *Ibid.* [25], p. 211.

En el contexto de la glutinometría al interior de la palabra, tenemos una serie de medidas para cada afijo. Es decir, cada segmento tiene una medida de afijalidad (o glutinosidad) en cada vocablo en el que aparece. Y, mientras más sean los vocablos en los que aparezca un afijo[65], mayor será el número de mediciones posibles $m_i(\dot{r})$ de su afijalidad $\dot{r}$ en el corpus y presuntamente más exacto será el valor $\bar{m}(\dot{r})$ de esta propiedad para el tipo abstracto que representa a cada ocurrencia (*token*) de dicho afijo. En otras palabras, el valor estimado de la afijalidad de cada afijo-tipo (miembro de un catálogo construido automáticamente) será el promedio de los valores calculados para cada ocurrencia del afijo en cada vocablo[66].

En cuanto al cálculo del error o desviación estándar σ para cada valor asignado a cada afijo-tipo, es necesario primero corroborar que las medidas obtenidas en cada vocablo en que ocurre el afijo en cuestión se distribuyan normalmente, en cuyo caso el error promedio $\bar{\varepsilon}$ se acercará a cero. Esa es, después de todo, la suposición que nos permite presumir que la afijalidad (o glutinosidad) de cada afijo-tipo converja a un valor real y, por lo tanto, sea susceptible de ser medida.

Por otra parte, un procedimiento similar serviría para estimar los errores de medición de la glutinosidad al exterior de la palabra. En principio, se puede calcular la glutinosidad (o cliticidad) inherente a uno de los extremos de un vocablo en diferentes córpora para obtener un conjunto de mediciones $m_i(\dot{r})$ que resulten en un promedio $\bar{m}(\dot{r})$ que represente la glutinosidad del extremo en cuestión del vocablo examinado. O se puede llevar a cabo el

[65]Nótese que no es lo mismo el número de vocablos en que aparece un afijo que su frecuencia en el corpus, ya que, como hemos visto, un segmento puede ocurrir en poquísimos vocablos de frecuencia considerable.

[66]Recuérdese que el índice de afijalidad calculado en el experimento del primer capítulo para cada afijo-tipo era el resultado de la asociación de los promedios de entropía, economía y cuadros. La idea ahora es calcular la afijalidad de cada ocurrencia de afijo en cada vocablo y luego promediarla para obtener la afijalidad del afijo-tipo que es miembro del catálogo.

mismo procedimiento no utilizando diferentes córpora, sino diferentes secciones del mismo corpus (por ejemplo, el *CEMC* está dividido en géneros)[67]. De esta manera, se obtendría para cada vocablo una estimación de un tipo de error involucrado en la medición de las glutinosidades asociadas a dicho vocablo. Aunque aquí el cálculo de este tipo de error no parece tan interesante como al interior de la palabra, es indudable que las consecuencias son muchas y valdría la pena investigarlas.

Por último, es necesario señalar que todo esto no agota las posibilidades de investigar al fenómeno del error. Por ejemplo, aquí falta considerar la relación de este fenómeno en el marco del estudio de la entropía, donde necesariamente el error o ruido se confunde con la incertidumbre o contenido de información del mensaje. Como quedó establecido en los capítulos anteriores, *información* en el sentido técnico es la medida de la libertad de opción al determinar un mensaje recibido a partir de varios posibles. Mientras mayor es la libertad de opción (es decir, a mayor información), mayor es la incertidumbre de que el mensaje seleccionado sea el correcto. Así, mayor libertad de decisión implica mayor incertidumbre y, por lo tanto, mayor información. Sin embargo, al introducir *ruido* en esta dinámica, podemos presumir que el mensaje recibido está de alguna manera distorsionado, es decir, hay material ajeno al mensaje o errores de algún tipo, que incrementan la incertidumbre de haber escogido el mensaje apropiado. Pero decir que la incertidumbre aumenta equivale a decir que el contenido de información transmitida aumenta también. Por eso, la incertidumbre causada por la libertad de escoger es deseable, mientras que aquella que resulta de los errores es

[67]Sin embargo, creo que antes de recurrir a tales complicaciones para afinar las estimaciones de glutinosidad de los vocablos, parece más prometedor tomar en cuenta contextos más amplios para cada vocablo (algo mayor a un digrama o digramas compuestos de dos secuencias de palabras gráficas).

indeseable[68]. Por lo pronto, queda la pregunta de si este error-incertidumbre es equiparable al ruido aleatorio de la teoría del error, o se trata de otra cosa que espera ser considerada para que podamos —como Varrón– seguirle la pista a los hechos lingüísticos con más fe de que no nos vamos a equivocar.

4.4 Glutinometría al interior del sintagma en el *CEMC*

En esta sección se presenta la aplicación del esquema elaborado en las secciones anteriores al *Corpus del Español Mexicano Contemporáneo*. Se examinan los resultados de los procedimientos para determinar las formas más gramaticales, las más clíticas y las hipotetizadas como nexos.

En la tabla 4.8 aparecen las cien formas más gramaticales del corpus (las primeras quinientas se listan en la tabla D.1 del apéndice), es decir, aquellas con los valores más altos de glutinosidad (a ambos lados)[69]. Según este esquema, el segmento más gramatical es la conjunción 'y', con un valor de poco más de 10.5 kiloVarrones. Los pronombres proclíticos exhiben cantidades que van entre los 3.39 kiloVarrones ('se') a 463 Varrones ('les'). Las preposiciones más importantes aparecen dentro de las 84 primeras formas (solamente faltan 'ante', 'contra' y 'bajo' que aparecen después de las cien primeras[70]).

Las conjunciones, así como todos los artículos y otros determinadores, se concentran

[68]Shannon y Weaver, *op. cit.* [125] 1964, p. 19.

[69]Se trata de glutinosidad escrita, es decir, la puntuación no se tomó en cuenta.

[70]Véase la tabla D.1 en el apéndice. Las formas 'según' y 'salvo' no aparecen dentro de las primeras quinientas, sino en los rangos 538 y 2052 (más allá de lo incluido en la tabla del apéndice). Sobra decir que el segmento 'cabe' (núm. 971) no representa la antigua preposición, sino al verbo 'caber'.

Tabla 4.8: Formas gramaticales del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
1.	y	60303	8638	1955	-6684	10590
2.	de	114346	7002	2260	-4743	9262
3.	en	50123	6963	238.3	-6725	7201
4.	del	18621	3805	1079	-2726	4884
5.	a	45525	4145	439	-3706	4584
6.	con	18747	3532	402.8	-3129	3935
7.	que	67243	2444	1135	-1309	3579
8.	se	33623	743.7	2649	1905	3393
9.	para	14655	2639	689.9	-1949	3328
10.	el	51708	1330	1977	647.5	3307
11.	por	19835	2998	213.9	-2784	3212
12.	al	11179	1996	1177	-819.2	3173
13.	la	73110	700.6	2315	1615	3016
14.	un	19765	522.9	2045	1522	2568
15.	los	31231	311.9	2020	1708	2331
16.	una	16473	457.9	1857	1399	2315
17.	su	12520	206.7	1934	1728	2141
18.	o	8264	1583	341	-1242	1924
19.	las	20882	258.3	1324	1065	1582
20.	como	11088	1011	363.8	-646.7	1374
21.	más	9778	501.9	869.6	367.7	1371
22.	sus	5267	80.28	1153	1073	1233
23.	me	10410	75.27	1149	1074	1224
24.	muy	4915	296.2	692.2	396	988.4
25.	es	18601	600.2	316	-284.2	916.3
26.	le	8502	56.63	846.9	790.3	903.6
27.	lo	13705	151.4	682.2	530.8	833.7
28.	nos	3399	60.45	718.9	658.5	779.4
29.	no	31676	350.6	421.9	71.28	772.5
30.	son	4447	290.8	404.2	113.4	695
31.	sin	3179	288.9	347	58.14	635.9
32.	sobre	2740	573.2	20.12	-553.1	593.3
33.	tan	1591	162	430.9	268.9	592.9

también entre los primeros segmentos. Además, ocurren de nuevo las formas verbales de los verbos con carácter más gramatical de la lengua española ('ser', 'estar', 'poder', 'haber', 'deber', 'tener', etc.).

Algunas formas verdaderamente léxicas (adjetivos como 'nacional', 'social', 'mexicana', 'internacional', sustantivos como 'vida', 'hombre', 'tiempo', 'trabajo' y adverbios como 'completamente', 'perfectamente', etc.) empiezan a ocurrir, espaciadamente, después de las

Tabla 4.8 (continuación):
Formas gramaticales del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
34.	te	3389	34.57	535.4	500.9	570
35.	entre	2528	534.1	33.57	-500.6	567.7
36.	dos	3216	49.92	497.4	447.5	547.3
37.	esta	2925	43.08	480.5	437.4	523.6
38.	está	4108	96.51	415.7	319.2	512.2
39.	ser	2830	23.34	450.4	427	473.7
40.	les	2021	26.67	436.3	409.6	463
41.	mi	4840	90.97	355.1	264.2	446.1
42.	cuando	4765	240.6	188.6	-52.02	429.2
43.	qué	4523	14.82	402.9	388.1	417.7
44.	e	1325	156.5	258	101.5	414.5
45.	esa	1666	37.36	370.1	332.7	407.4
46.	puede	2480	53.11	350.3	297.1	403.4
47.	están	1561	54.81	338.8	284	393.6
48.	también	3501	259.4	134	-125.5	393.4
49.	hasta	3018	316.5	55.95	-260.6	372.5
50.	porque	5087	259.8	100.9	-158.9	360.6
51.	había	2023	26.25	328.5	302.3	354.8
52.	ya	9791	202.8	150.6	-52.11	353.4
53.	ese	1977	34.07	312.9	278.9	347
54.	pero	8336	230	107.7	-122.3	337.6
55.	gran	1182	5.579	317.8	312.2	323.4
56.	ha	4105	44.1	277.4	233.3	321.5
57.	fue	2827	138.3	171.3	32.96	309.6
58.	mucho	1754	245.8	62.16	-183.6	307.9
59.	así	4161	248.4	47.97	-200.4	296.4
60.	tu	1278	41.72	250.8	209.1	292.5
61.	bien	2603	98.52	190.8	92.28	289.3
62.	yo	7044	157.3	129.2	-28.07	286.5
63.	han	1947	27.36	257.9	230.6	285.3
64.	ni	2392	179.7	103.4	-76.21	283.1
65.	si	6122	154.1	127.1	-26.98	281.3
66.	debe	1318	70.18	183.2	113	253.4

primeras cien (tabla D.1 del apéndice). En contraste, segmentos poco frecuentes pero de indudable carácter gramatical tienen rangos similares (núm. 133, 'conmigo'; núm. 145. 'cualquier'; núm. 105, 'cuya'; y núm. 265, 'cuyo').

Es de notarse que el índice de ordenamiento (la suma de las glutinosidades de los lados) disminuye muy rápidamente, de tal manera que entre la primera y la forma número cien hay alrededor de 9 kiloVarrones de diferencia, mientras que ésta última ('hacer') apenas alcanza

Tabla 4.8 (continuación):
Formas gramaticales del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
67.	hacia	911	241.5	10.96	-230.5	252.5
68.	aquí	4024	180.7	69.89	-110.9	250.6
69.	otros	1499	9.734	240	230.3	249.8
70.	99	16449	108.6	138.2	29.61	246.9
71.	pueden	944	36.05	204.1	168.1	240.2
72.	sí	9079	183.5	51.22	-132.3	234.7
73.	tiene	3122	73.98	155	81.03	229
74.	tres	1598	32.81	195.2	162.4	228
75.	hay	4013	34.02	193.2	159.2	227.3
76.	grandes	845	31.76	192.1	160.3	223.9
77.	estas	765	12.47	211.1	198.7	223.6
78.	nada	2730	184.3	37.75	-146.5	222
79.	era	2650	84.56	130.6	46.01	215.1
80.	sido	1050	0.6681	213.4	212.8	214.1
81.	mis	963	26.41	181.3	154.8	207.7
82.	estos	1005	9.587	192.1	182.5	201.6
83.	otro	1979	30.02	168.9	138.9	198.9
84.	desde	1904	174.7	23.17	-151.6	197.9
85.	esos	741	6.908	186.9	180	193.8
86.	fueron	861	64.53	128.1	63.58	192.6
87.	nuestra	768	10.8	180.3	169.5	191.1
88.	estaba	1368	37.62	151.6	114	189.2
89.	otras	1071	12.41	171.4	159	183.8
90.	uno	3075	127.8	53.38	-74.42	181.2
91.	mal	579	84.73	96.23	11.51	181
92.	donde	1968	49	131.1	82.07	180.1
93.	bastante	510	70	106.3	36.27	176.3
94.	ahí	2060	88.07	84.46	-3.601	172.5
95.	muchos	941	26.47	144	117.5	170.4
96.	pues	6077	91.09	77.68	-13.4	168.8
97.	entonces	2974	88.77	79.99	-8.782	168.8
98.	unas	643	42.17	126.2	84.03	168.4
99.	mejor	1110	41.03	127.2	86.21	168.3
100.	hacer	1707	37.57	128.6	91.07	166.2

166.2 Varrones. De hecho, la número 500 ('distintas') tiene un valor de 42.53 Varrones.

En la tabla 4.9 se consignan los segmentos cuya diferencia entre las glutinosidades de cada extremo es mayor. El índice de ordenamiento es el mismo utilizado en el capítulo anterior (la diferencia multiplicada por la menor de las cliticidades o glutinosidades). Los valores más altos del índice de ordenamiento (Varrones^2) y, en gran medida, de las diferencias absolutas corresponden a las formas con menos probabilidad de aparecer solas (todos los pronombres

Tabla 4.9: Proclíticos del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
1.	se	33623	743.8	2649	1905	5047000
2.	la	73110	700.5	2316	1615	3740000
3.	los	31231	311.9	2019	1707	3448000
4.	su	12520	206.7	1934	1728	3342000
5.	un	19765	522.9	2045	1522	3113000
6.	una	16473	457.9	1857	1399	2598000
7.	las	20882	258.3	1324	1066	1411000
8.	el	51708	1330	1977	647.7	1281000
9.	sus	5267	80.27	1153	1073	1237000
10.	me	10410	75.27	1149	1074	1234000
11.	le	8502	56.63	846.9	790.3	669300
12.	nos	3399	60.45	718.9	658.5	473400
13.	lo	13705	151.4	682.2	530.8	362100
14.	más	9778	501.9	869.5	367.7	319700
15.	muy	4915	296.2	692.2	396	274100
16.	te	3389	34.57	535.4	500.9	268200
17.	dos	3216	49.92	497.4	447.5	222600
18.	esta	2925	43.08	480.5	437.4	210200
19.	ser	2830	23.34	450.4	427	192300
20.	les	2021	26.67	436.3	409.6	178700
21.	qué	4523	14.82	402.9	388.1	156400
22.	está	4108	96.51	415.7	319.2	132700
23.	esa	1666	37.36	370.1	332.7	123100
24.	tan	1591	162	430.9	268.9	115900
25.	puede	2480	53.11	350.3	297.1	104100
26.	había	2023	26.25	328.5	302.3	99300
27.	gran	1182	5.579	317.8	312.2	99220
28.	están	1561	54.81	338.8	284	96230
29.	mi	4840	90.97	355.1	264.2	93820
30.	ese	1977	34.07	312.9	278.9	87260
31.	ha	4105	44.1	277.4	233.3	64740
32.	han	1947	27.36	257.9	230.6	59480
33.	otros	1499	9.734	240	230.3	55280

clíticos, todos los artículos, adjetivos posesivos (su, sus, mi, tu, etc.), prenominales (gran, dos, etc.) y determinadores de otros tipos (esta, esa, ese, etc.), adverbios (muy, tan, etc.). Es obvio que al decrecer el índice de ordenamiento, el carácter prototípico de los segmentos como clíticos disminuye rápidamente (los clíticos son pocos) de tal manera que los segundos cincuenta segmentos, aunque se trate de formas muy gramaticales, están mucho más alejados del clítico prototípico que los primeros 25 segmentos. De nuevo, me parece que la ocurrencia de formas verbales es muy interesante. Son unas pocas formas después de las cuales suceden

Tabla 4.9 (continuación):
Proclíticos del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
34.	tu	1278	41.72	250.8	209.1	52440
35.	son	4447	290.8	404.2	113.4	45850
36.	sido	1050	0.6681	213.4	212.8	45410
37.	estas	765	12.47	211.1	198.7	41950
38.	estos	1005	9.587	192.1	182.5	35050
39.	pueden	944	36.05	204.1	168.1	34300
40.	esos	741	6.908	186.9	180	33650
41.	tres	1598	32.81	195.2	162.4	31700
42.	grandes	845	31.76	192.1	160.3	30800
43.	hay	4013	34.02	193.2	159.2	30760
44.	nuestra	768	10.8	180.3	169.5	30550
45.	no	31676	350.6	421.9	71.27	30070
46.	mis	963	26.41	181.3	154.8	28070
47.	otras	1071	12.41	171.4	159	27240
48.	e	1325	156.5	258	101.5	26190
49.	otro	1979	30.02	168.9	138.9	23450
50.	misma	887	2.154	152.2	150.1	22840
51.	cuya	267	8.143	153.2	145	22210
52.	debe	1318	70.18	183.2	113	20700
53.	buen	436	8.521	147.8	139.3	20580
54.	sin	3179	288.9	347	58.14	20170
55.	mismo	1801	4.259	139.7	135.4	18920
56.	nueva	452	4.336	135.8	131.5	17860
57.	habían	443	13.06	140.1	127.1	17810
58.	bien	2603	98.52	190.8	92.28	17610
59.	estaba	1368	37.62	151.6	114	17280
60.	muchos	941	26.47	144	117.5	16910
61.	quien	998	12.98	135.3	122.3	16550
62.	haber	604	12.98	131.3	118.3	15540
63.	estar	665	30.25	131.3	101.1	13280
64.	hacen	733	9.439	119.9	110.5	13250
65.	nuestros	445	7.278	118.3	111	13130
66.	mayor	1209	32.81	132.1	99.29	13120

otras muchas formas. Así, es sumamente probable, como se señaló en el capítulo anterior, que ocurra algo después de 'ser', 'está', 'puede', 'debe', etc. (formas que pueden ir seguidas sobre todo de participios, gerundios o infinitivos) y 'hay' (muy seguida de sustantivos).

Por último, la tabla 4.10 contiene las formas del corpus cuyas glutinosidades de cada lado son similares. El índice de importancia es el cociente de la menor glutinosidad sobre la mayor divido entre su diferencia relativa (su diferencia absoluta en Varrones dividida entre la menor

Tabla 4.9 (continuación):
Proclíticos del *CEMC*

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
67.	estoy	693	20.11	123.8	103.7	12840
68.	tiene	3122	73.98	155	81.03	12560
69.	diferentes	412	19.48	122.1	102.7	12540
70.	vida	1632	1.448	110.9	109.5	12150
71.	hace	2003	20.8	120.8	99.96	12070
72.	hacer	1707	37.57	128.6	91.07	11710
73.	cualquier	655	8.051	111.4	103.4	11520
74.	esas	658	11.68	113.1	101.4	11470
75.	puedo	439	7.498	109.1	101.6	11090
76.	algunos	822	36.93	124.9	87.95	10980
77.	mejor	1110	41.03	127.2	86.21	10970
78.	donde	1968	49	131.1	82.07	10760
79.	nuevo	544	6.652	107.1	100.4	10750
80.	unas	643	42.17	126.2	84.03	10610
81.	cuatro	833	23.27	114.7	91.43	10490
82.	primera	811	0.8291	100.4	99.58	10000
83.	nuestro	802	8.262	104	95.72	9952
84.	tienen	1275	25.73	112.8	87.09	9826
85.	cada	1878	23.91	111.5	87.64	9775
86.	solo	405	18.03	108.1	90.05	9732
87.	varios	438	23.53	110.8	87.29	9673
88.	este	7157	39.48	119.9	80.44	9646
89.	buena	519	15.81	106.4	90.62	9644
90.	poder	514	8.284	102.4	94.11	9637
91.	estaban	402	28.05	111.9	83.85	9383
92.	nuestras	283	4.929	99.05	94.12	9322
93.	forma	1495	4.904	96.87	91.97	8909
94.	primer	788	1.671	95.12	93.45	8889
95.	pueda	320	9.863	97.47	87.61	8540
96.	fueron	861	64.53	128.1	63.58	8145
97.	demás	483	0.2108	89.86	89.65	8056
98.	dar	755	11.33	95.41	84.08	8023
99.	él	2633	19.16	96.19	77.03	7409
100.	menos	1413	13.88	93.02	79.13	7361

de las glutinosidades): $\frac{min(GL(s_x))^2}{max(GL(s_x))\times[max(GL(s_x))-min(GL(s_x))]}$. Se escogió este índice, porque el utilizado en el capítulo anterior no prevenía que ocurrieran en la lista segmentos con una diferencia absoluta de varios cientos o, incluso, miles de Varrones. Por ejemplo, el segmento 'y' que en el capítulo pasado exhibió una diferencia normalizada de -0.1037 (véase la tabla 3.8 de la página 216) tiene, de hecho, una gigantesca diferencia absoluta de -6.684 kiloVarrones. Esta importante cantidad se oculta al considerarse solamente su dimensión relativa. Si de lo

que se trata es de identificar las formas más gramaticales con glutinosidades semejantes, vale la pena considerar las diferencias absolutas menores.

Tabla 4.10: Formas del *CEMC* con cliticidades cercanas

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
1.	será	747	76.75	76.67	-0.08411	910.5
2.	muchas	851	42.73	42.79	0.05892	724.2
3.	feliz	162	31.41	30.79	-0.6157	49.03
4.	viene	555	32.44	33.33	0.8944	35.29
5.	ahí	2060	88.07	84.46	-3.601	22.5
6.	ahora	1923	58.99	62.94	3.956	13.97
7.	allá	1229	59.95	56.1	-3.848	13.64
8.	únicamente	216	28.97	31.13	2.155	12.52
9.	usté	728	63.37	58.67	-4.708	11.53
10.	casi	1133	61.05	66.82	5.772	9.664
11.	entonces	2974	88.77	79.99	-8.782	8.207
12.	orita	533	39.23	44	4.765	7.341
13.	mal	579	84.73	96.23	11.51	6.483
14.	cómo	1694	56.49	64.22	7.733	6.425
15.	pues	6077	91.09	77.68	-13.4	4.944
16.	esto	1542	41.21	48.76	7.553	4.611
17.	pasar	373	30.31	25.35	-4.962	4.272
18.	sin	3179	288.9	347	58.14	4.136
19.	no	31676	350.6	421.9	71.28	4.088
20.	si	6122	154.1	127.1	-26.98	3.887
21.	yo	7044	157.3	129.2	-28.07	3.781
22.	nomás	813	51.59	63.78	12.19	3.424
23.	fue	2827	138.3	171.3	32.96	3.389
24.	va	1907	31.2	38.82	7.622	3.289
25.	verde	156	24.37	30.55	6.175	3.149
26.	allí	973	67.41	85.08	17.66	3.025
27.	pequeños	182	23.28	29.44	6.154	2.993
28.	principal	258	31.66	40.28	8.623	2.886
29.	99	16449	108.6	138.2	29.61	2.883
30.	todo	4119	63.32	49.66	-13.65	2.853
31.	cuando	4765	240.6	188.6	-52.02	2.842
32.	todavía	780	65.6	51.25	-14.35	2.791
33.	usted	1482	65.33	83.85	18.52	2.748

De esta manera, en la tabla 4.10 se listan las formas con más de 50 Varrones de glutinosidad total (de un lado sumada a la del otro), cuya diferencia es menor según el índice de ordenamiento descrito arriba. Además, se requirió que el tamaño de la diferencia fuera menor a la menor de las glutinosidades. El resultado está, como se dijo, en la tabla 4.10. Las formas no alcanzan a ser, como se ve, ni siquiera noventa y la diferencia absoluta no deja

Tabla 4.10 (continuación):
Formas del *CEMC* con cliticidades cercanas

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
34.	escrito	134	29.46	22.69	-6.774	2.579
35.	vienen	352	30.73	23.12	-7.613	2.284
36.	totalmente	184	35.16	26.42	-8.741	2.271
37.	anterior	460	30.92	22.98	-7.941	2.15
38.	ya	9791	202.8	150.6	-52.11	2.148
39.	nunca	1046	47.65	65.63	17.98	1.923
40.	anda	217	25.36	34.96	9.608	1.914
41.	dinero	615	46.38	33.41	-12.97	1.857
42.	son	4447	290.8	404.2	113.4	1.844
43.	política	476	26.89	38.16	11.27	1.682
44.	mayores	237	53.8	37.52	-16.28	1.608
45.	completamente	191	44.16	63.86	19.7	1.55
46.	actual	285	31.82	21.77	-10.06	1.48
47.	hacerlo	214	20.97	30.81	9.84	1.45
48.	bueno	2328	31.43	21.33	-10.1	1.433
49.	el	51708	1330	1977	647.5	1.381
50.	bastante	510	70	106.3	36.27	1.271
51.	era	2650	84.56	130.6	46.01	1.19
52.	demasiado	214	40.65	63.06	22.41	1.17
53.	serán	209	28.99	45.06	16.07	1.161
54.	hoy	743	30.66	48.02	17.37	1.127
55.	tanto	1418	29.63	47.87	18.25	1.005
56.	estuvo	378	34.2	55.31	21.11	1.001
57.	e	1325	156.5	258	101.5	0.9347
58.	fuera	736	41.13	24.93	-16.2	0.9331
59.	mexicanos	339	46.47	28.02	-18.45	0.9158
60.	grande	613	33.23	19.77	-13.45	0.8746
61.	al	11179	1996	1177	-819.2	0.8468
62.	siempre	1655	59.58	101.4	41.85	0.8363
63.	cinco	750	23.25	39.94	16.69	0.8107
64.	más	9778	501.9	869.6	367.7	0.7878
65.	ni	2392	179.7	103.4	-76.21	0.7815
66.	perfectamente	136	40.31	22.88	-17.43	0.7449

de ser considerable al compararse con las mediciones de glutinosidad. De hecho, podemos ver que la condición de "nexos" que se esperaba en el capítulo pasado que tuvieran ciertos segmentos fue tal vez una intuición precipitada. Si bien ocurren algunas de las formas que se habían considerado como tales (conjunciones como 'e', 'ni' —núms. 57 y 65— y formas del verbo ser como 'será', 'fue', 'son', etc.), la mayoría —sobre todo al principio de la tabla— no son de ese tipo. Quizá la intuición de que se trata de las formas más independientes (es decir, que funcionan como sintagmas o unidades completas en relativamente cualquier

Tabla 4.10 (continuación):
Formas del *CEMC* con cliticidades cercanas

	clítico	fr.	cliticidad de un lado	cliticidad del otro	diferencia	índice de ordenamiento
67.	ahorita	224	19.29	34.04	14.75	0.7414
68.	andan	104	33.04	18.49	-14.55	0.7108
69.	tenían	267	21.12	38.11	16.98	0.6897
70.	todos	2627	40.6	22.42	-18.18	0.6811
71.	grupos	280	18.34	33.35	15	0.6727
72.	realmente	242	18.74	34.17	15.42	0.6665
73.	acá	547	21.34	38.97	17.64	0.6624
74.	ir	824	51.68	28.15	-23.53	0.6514
75.	solamente	460	19.55	35.9	16.35	0.6512
76.	otra	2082	50.81	93.55	42.74	0.6457
77.	mexicano	391	53.23	28.44	-24.79	0.6128
78.	es	18601	600.2	316	-284.2	0.5856
79.	unos	1166	55.06	104.8	49.72	0.582
80.	eran	561	51.24	97.56	46.32	0.581
81.	también	3501	259.4	134	-125.5	0.5513
82.	bien	2603	98.52	190.8	92.28	0.5513
83.	tú	1297	24.62	48.33	23.71	0.5288
84.	fueron	861	64.53	128.1	63.58	0.5112

posición de la oración) sea más acertada. Pero son muy pocas formas para corroborar esto (y la mayoría tiene una glutinosidad total más bien baja). Lo único que queda claro es que todavía falta investigar en otros córpora y para otras lenguas la naturaleza de los segmentos con glutinosidades similares a cada lado.

El caso es que a partir de los datos en estas tablas se puede observar que hay por lo menos cuatro tipos de segmentos: aquellos muy gramaticales con una gran diferencia entre glutinosidades, aquellos muy gramaticales con una mínima diferencia entre glutinosidades, aquellos con poca glutinosidad total (poco gramaticales) con una diferencia importante y aquellos poco gramaticales con una mínima diferencia entre glutinosidades. A grandes rasgos, los primeros corresponderían a los clíticos prototípicos. Los segundos a otros tipos de palabras función. Los terceros a ciertas palabras de contenido que tienden a adherirse a otras (como los adjetivos y los adverbios). Finalmente, los últimos corresponderían a las palabras propiamente léxicas, como los nombres y los verbos, es decir, todos los segmentos que sue-

len transmitir el contenido propio de los textos, que contienen un mínimo de información gramatical y que no se adhieren a los segmentos que los rodean.

4.5 Los signos gramaticales del español de México

En esta sección se analiza el carácter del conjunto de signos presentado en el apartado anterior como de mayor uso gramatical del corpus. El conjunto se muestra ampliado en la tabla D.1 del apéndice. donde se consignan 500 formas. La idea es examinar los usos típicos de estos segmentos para determinar qué tan gramatical es su naturaleza. es decir. si se emplean como palabras función o de contenido.

Sin duda, no es cosa trivial caracterizar la noción de lo gramatical con criterios más complejos que el de la simple frecuencia. Por eso, hasta aquí se ha trabajado con la idea de *uso* gramatical, que parece más apropiado para referirse a los asuntos cuantitativos de lo gramatical. Sin embargo, vale la pena, aunque sea brevemente, considerar algunos aspectos cualitativos de este fenómeno.

Si bien se ha escrito mucho sobre 'gramaticalización', aquí me referiré solamente a algunas cuestiones de esa discusión con el objeto de examinar los resultados de todo lo propuesto en este trabajo. Típicamente. el término gramaticalización se emplea en contextos de cambio diacrónico. Este trabajo. sin embargo, se ha enfocado principalmente al aspecto sincrónico. por lo que mucho de lo que se discute alrededor de ese concepto no se tocará aquí[71].

[71] El cambio ocurre en contextos muy locales; se dispara por factores pragmáticos, lo que significa que no es arbitrario del todo; etc.

Una cuestión central es que los signos gramaticales se originan de formas relativamente vacías de significado. Esto no quiere decir que no signifiquen nada, sino que en detrimento de significados concretos, sus significados tiendan a ser relativamente abstractos. De hecho, son las formas de ciertos lexemas (o clases de lexemas), y no de otros, las que tienen más posibilidades de ser utilizadas para codificar categorías gramaticales. Por ejemplo, vocablos tales como 'piñata' o 'confesar' seguramente tendrían que sufrir varios cambios semánticos muy importantes antes de *usarse* como palabras gramaticales, porque se refieren a cosas muy específicas y, al tener una frecuencia relativamente baja, difícilmente se utilizan en contextos restringidos sintácticamente (aunque ocurran en tipos de textos específicos). Esto significa que los lexemas más aptos de gramaticalizarse pertenecen a un grupo limitado de campos léxicos. Y, dentro de estos campos, aquellos de naturaleza superordinada tienen más posibilidades de fungir como signos gramaticales. Por ejemplo, 'cosa' y 'objeto' son formas más aptas que 'piñata' de convertirse en palabras gramaticales, porque son sus hiperónimos.

Los términos en camino de convertirse en signos gramaticales tienen típicamente diferentes lecturas en sus diferentes contextos. Esto es, producen diversas inferencias, según sus entornos, en lo que se refiere a sus significados. En unos la lectura puede ser más gramatical, en otros más de contenido. Por ejemplo, 'acuerdo' funciona de manera distinta en `'el acuerdo del partido'` que en `'de acuerdo con el partido'`. En la primera es un sustantivo, mientras que en la segunda no lo es tan plenamente.

Esto implica que las formas *viejas* coexisten con formas nuevas. Pero esta convivencia apoya o refuerza los significados viejos en las formas nuevas. Esto es, los significados de formas en diferentes estados de gramaticalización persisten debido a las lecturas de las formas más

antiguas.

El que un signo ocurra como gramatical o de contenido en diferentes situaciones lingüísticas, significa que la manera en que se analiza es diferente en esos contextos. Típicamente, una construcción se analiza en algún lugar como un grupo de cosas, mientras que en otro la construcción se analiza como una unidad. Por eso, en el ejemplo anterior la expresión 'de acuerdo con' significa 'según'. Pero, en referencia a la persistencia de significados, conserva el sentido de 'convenio'. Además, parece común que allí donde se advierte este tipo de reanálisis, también se observa algún tipo de reducción fonológica, porque la expresión ya no es analizable en sus partes.

Con esto en mente, es de esperarse que entre los segmentos que se aislaron como resultado de este experimento se exhiban algunas de estas características. Por ejemplo, debido al desgaste fonológico y semántico esperado en vocablos muy usados gramaticalmente, no sorprende que 303 de las formas midan cinco fonemas o menos, 192 tienen cuatro o menos, 97 tienen tres o menos y 44 tienen menos de tres. Esto es cierto no sólo al medir el tamaño de los vocablos en número de fonemas, sino también en cantidad de morfemas: mientras menos material fonológico, menos espacio para morfemas (en lenguas aglutinantes habría al menos un fonema por cada morfema).

Así, dentro de las primeras treinta formas de las tablas 4.8 (presentada arriba) y D.1 (del apéndice), la mayoría no son analizables en morfemas. Luego, al examinar el resto nos encontramos con más formas analizables en elementos morfológicos. De hecho, mientras que aquellas con rango mayor se prestan a análisis sistemáticos a los que llegarían hablantes promedio (por ej., 'distintas' —rango 500— se analizaría en la raíz *distint~* y los sufijos de

flexión $\sim a$ y $\sim s$), aquellas con rango menor no se analizan en partes menores ('y', 'de', 'en', 'muy', 'cuando', etc.), o ninguna de sus partes constituye una verdadera raíz ('de-l' y 'a-l': 'l-o-s', 'l-a-s', 'su-s', 'es-a', etc.), o su significado conjunto no es el de la suma de sus partes ('tam-bién' y 'por-que').

Lo interesante es que, como los sufijos, muchas formas constituyen paradigmas. Algunos son paradigmas con formas predominantemente supletivas, especialmente cuando no son analizables ('yo', 'tú', 'él', etc.) o dejaron de serlo ('conmigo', 'tampoco'). Pero la gran mayoría exhibe estructuras analizables y, por lo tanto, los morfemas en que se analizan sirven para agruparlas en paradigmas nominales, verbales y adverbiales. De hecho, varios de esos paradigmas tienen la misma base, lo que indica la posibilidad de lematizarlas, es decir, encontrar una forma canónica que las represente.

La cuestión es que, analizables o no, las formas se pueden agrupar de diversas maneras. Una sería por categoría gramatical. Por ejemplo, para construir una jerarquía de categorías, se puede calcular el promedio de rangos por categoría. En la tabla 4.11 se ordenan las categorías de menor a mayor promedio.

Tabla 4.11: Promedios de rangos por categorías gramaticales

categoría	promedio de rangos
conj.	36
art.	39
prep.	61
pron.	148
det.	159
adv.	227
v.	260
adj.	307
sust.	360

Aquí observamos una progresión que va de signos muy cortos y bien definidos, como partes

invariables de la oración (conjunciones, artículos y preposiciones), a signos más complejos de clases abiertas (adverbios, verbos, adjetivos y sustantivos).

Una manera de presentar estos paradigmas es agrupándolos según las posiciones típicas en las que ocurren al interior de (o con respecto a) sintagmas nominales y verbales. En la tabla 4.12 se presenta un esquema muy simplificado de los lugares específicos dentro de los sintagmas nominales (o preposicionales) en que típicamente ocurren las formas más gramaticales de la lista del apéndice.

Tabla 4.12: Esquema simplificado de paradigmas de formas gramaticales (el sintagma nominal)

SP					
	SN				
	pron.				
prep.	**det.**[a]	**adj.**	**núcleo**	**adv.**[b]	**adjs.**
	del	gran		ya	
	al	buen		no	
a	el, la, lo	mal		tan	
de	ese, esa	primer		muy	
en	este, esta	[tercer]		más	
con	mi, tu, su, nuestro, nuestra			bien	
por	aquel, aquella			casi	~o
sin	un, una, algún, alguna		**sust. sing.**	algo	~a
ante	ningún, ninguna			poco	~e
bajo	otro, otra			menos	
entre	tanto, [tanta,] tal			nunca	
sobre	cada, cual, cualquier			siempre	
hasta	todo (el/ese/este/un/mi...)			también	
hacia	toda (la/esa/esta/una/mi...)			bastante	
desde				demasiado	
contra				~mente	
durante	algo, nada, nadie, [alguien, ninguno,] ninguna, lo demás,			·	
para/pa	él, ella, [ello, ése, ésa,] eso, éste, [ésta,] esto,			·	
[según]	uno, una, todo, toda, alguno, alguna, [aquello,] aquella,			·	
[tras]	otro, otra, tanto[, tanta, cualquiera]...				

[a]Las formas no atestiguadas dentro de las 500 más gramaticales (véase la tabla D.1) aparecen entre corchetes cuadrados.

[b]Más adelante se examinan los adverbios.

Las tabla habla por sí misma. En la primera parte aparecen los nichos correspondientes

a los sintagmas de sustantivos en singular y, en la segunda, que se encuentra una página después, se exhiben los correspondientes a los sustantivos en plural. Como antes, las formas entre corchetes no están atestiguadas en la tabla D.1 del apéndice. Aunque no estén todas, se ve que hay una muestra muy rica de palabras función del sintagma nominal en la lista del apéndice. De todas éstas, solamente las distributivas 'sendos' y 'sendas' no ocurrieron ni siquiera dentro de las 6000 de más uso gramatical ni dentro de las 6000 más frecuentes.

Tabla 4.12 (continuación):
Esquema simplificado de paradigmas de formas gramaticales (el sintagma nominal)

SP					
	SN				
	pron.				
prep.[a]	**det.**	**adj.**	**núcleo**	**adv.**[b]	**adjs.**
	los, las	dos			
	esos, esas	tres			
	estos, estas	cuatro			
	mis, tus, sus, nuestros, nuestras	cinco			
	[aquellos, aquellas]	seis			~os
	unos, unas, algunos, algunas	[...]	**sust. pl.**		~as
	otros, otras	diez			~es
	[tantos, tantas,] tales	[...]			
	ambos[, ambas, sendos, sendas]	mil			
	todos (los/esos/estos/un/mis...)	[...]			
	[todas (las/esas/estas/un/mis...)]	99[c]			
	ellos, ellas, unos, unas, [ésos, ésas, éstos, éstas,] todos[, todas, aquellos, aquellas, tantos, tantas,] algunos, algunas, otros, otras, ambas...				

[a]Véase la lista de preposiciones del principio de la tabla.
[b]Véase la lista de adverbios del principio de la tabla. Más adelante se examinan con detalle.
[c]Como se mencionó en el primer capítulo. '99' representa cualquier combinación de dígitos.

Una diferencia interesante entre la primera parte y la segunda es el tipo de formas adjetivas que sólo ocurren antes del sustantivo. En el sintagma nominal plural son típicamente numerales. Es muy interesante que entre las formas del apéndice haya ocurrido aquella que representa cualquier combinación de dígitos. Por supuesto, el procedimiento aplicado no rinde cuenta de los contextos en los que ocurre. Pero, al examinar el corpus, se constata que la mayoría de las ocurrencias de series consecutivas de dígitos aparecen asociadas a un sus-

tantivo (por ejemplo, alguna moneda, términos de fórmulas, etc.), expresando la cantidad de aquello que designa. Difícilmente podríamos considerar cualquier secuencia de dígitos como gramatical, pero estas secuencias ocurren en posición definitivamente gramatical. Dada la variedad de números en el corpus, aquí se observa claramente que lo gramatical es más que el signo mismo. En otras palabras, muchísimo de su carácter depende de su posición en el sintagma en relación con los otros signos. De hecho, el signo gramatical no lo es sin su contexto.

Si bien, por un lado, esta tabla no es de ninguna manera el mejor esquema para representar al sintagma nominal (ni al preposicional), por el otro, hace evidente que se trata de una estructura con lugares específicos para cada cosa: las preposiciones no van después de los adjetivos demostrativos, los artículos no van después del sustantivo, los adverbios tienen su lugar antes de los adjetivos. Los adjetivos representados por la última columna (y que son clase abierta) bien pueden colocarse antes del sustantivo, pero en general la mayoría de las otras formas sencillamente no ocurren en otros lugares, ni siquiera por razones estilísticas.

De todos los nichos del esquema de la tabla 4.12, lo único verdaderamente libre en cuanto a su ocurrencia en la oración es el núcleo sustantivo. En cierta manera, también ese es el caso de las formas pronominales, porque funcionan por sí mismas como núcleos del sintagma. Es curioso que las formas "faltantes" sean todas pronombres, aunque algunas sean además adjetivos. Con excepción de 'sendos' y 'sendas', los determinativos que no están en la lista del apéndice exhiben las mismas formas que los pronombres, es decir, tienen diferentes lecturas en diferentes contextos, unas más de uso gramatical (como determinativos), otras más de contenido (como núcleo de sintagma). De hecho, al tratarse de formas que ocurren tan a

menudo, resulta que por frecuencia algunos tienen rangos menores que en la parte de la tabla D.1 que por falta de espacio no se exhibe ni siquiera en el apéndice: por ejemplo, 'alguien' en la tabla del apéndice tiene rango 839 (30.1 varrones), pero por frecuencia tiene rango 750 (218 ocurrencias en el corpus). Casos similares son los de 'ése', 'ésa', 'ésos', 'ésas', 'éstos', 'éstas' que, al igual que 'alguien', fungen como núcleos de sintagma nominal. Si los percibimos como muy usados gramaticalmente, es por su valor anafórico, fenómeno que no se contempló en este trabajo pero que sería interesantísimo investigar en un futuro para determinar las maneras de cuantificarlo.

Por otra parte, las formas que, además de fungir como pronombres, se utilizan como adjetivos demostrativos obtienen mejores rangos al ordenarlas por el índice de varrones que por frecuencias, aunque no ocurran dentro de las 500 de la tabla del apéndice. Por ejemplo. 'tanta', 'aquellas' y 'ambas' por frecuencia tienen rangos altísimos (núms. 1576, 1013 y 1261 respectivamente), porque no son tan comunes como las formas masculinas. Sin embargo, ordenadas por cantidad de varrones, se acercan considerablemente a las 500 más usadas gramaticalmente (con rangos 639, 524 y 693 respectivamente).

Otro paradigma importante que no podía faltar es el de los pronombres personales que se exhiben en la tabla 4.13. De nuevo. aquellos no atestiguados están entre corchetes. Curiosamente, el caso de los posesivos es similar al de los pronombres del sintagma nominal. Con excepción de 'mía' y 'tuya', las formas posesivas faltantes obtienen mejores rangos ordenadas por frecuencia que por varrones. El pronombre 'mí' también. Sin embargo, 'consigo' y 'nosotras' muestran la situación contraria. Son relativamente poco frecuentes (44 y 33 respectivamente), por lo que por frecuencia obtienen rangos enormes (3857 y 4522). Sin

Tabla 4.13: Pronombres personales

clíticos	pron.	con prep.	adj. enfático	posesivos
me	yo	[mí] conmigo	mismo/misma	[mío, mía, míos, mías]
te	tú	ti contigo		[tuyo, tuya, tuyo, tuyas]
se	usted, usté	sí [consigo]		[suyo, suya, suyos, suyas]
lo, la le, se	él, ella			
nos	nosotros[, nosotras]		mismos/mismas	nuestro, nuestros, nuestra, nuestras
se	ustedes	sí [consigo]		[suyo, suya, suyos, suyas]
los, las les, se	ellos, ellas			

embargo, su posición mejora considerablemente al ordenarlas por varrones (2353 y 2047).

Lo importante de la tabla 4.13 es que muestra que los otros nichos de los paradigmas allí representados están ocupados. De hecho, formas tan gramaticales como 'conmigo', 'contigo' y 'ti' no son formas tan comunes, por lo que no estarían en esta tabla de haber considerado sólo sus frecuencias.

En el nivel de la oración encontramos que otras formas también tienen su lugar específico. El esquema de la tabla 4.14 las muestra en los lugares en que aparecerían en una oración española simplificada. Obviamente, este esquema no es muy fino, porque no se aprecia en él la riqueza de combinaciones posibles de formas en la oración española, pero sí muestra que hay una estructura y que la mayoría de los objetos que la conforman sí fueron aislados en la tabla del apéndice.

Entre las cosas que funcionan como marcas de subordinación hay varias cosas de naturaleza variada (adverbios, pronombres, adjetivos), que sirven para establecer relaciones hipotácticas entre oraciones. Típicamente ocurren entre las conjunciones, que establecen relaciones paratácticas y las estructuras oracionales, esto es, ocupan lugares muy específicos en relación con la oración. En general, las formas que no ocurrieron en la tabla del apéndice

Tabla 4.14: Esquema simplificado de paradigmas de formas gramaticales (la oración)

O					
				SV	
conj.	**subord.**	**...**	**adv.**	**pron.**	**núcleo**
y/e	si		no	se	
o[/u]	que		sí	me	
ni	como		~mente	te	
pero	cuando		·	nos	
	[cuanto]		·	le	
	porque		·	les	
	aunque			lo	v.
	donde, onde	...		los	
	quien, quienes			la	
	(el/lo/la/los/las) que			las	
	(el/lo/la) cual				
	(los/las) cuales				
	cuyo[, cuyos] + SN				
	cuya[, cuyas] + SN				
	qué, dónde, cómo, por qué				
	[quién, quiénes, cuándo, cuánto]				

tienen rangos no muy alejados de esas 500. Curiosamente, los pronombres interrogativos 'quién' y 'quiénes' exhiben un comportamiento similar al de los pronombres del sintagma nominal: sus rangos por frecuencia fueron menores que sus rangos por uso gramatical.

Por último, examinaremos los vocablos menos aptos de ser considerados gramaticales principalmente porque pertenecen a clases abiertas: adverbios, verbos, sustantivos y adjetivos, en ese orden, que es de más a menos gramatical, según los promedios de rangos en la tabla de los 500.

Un grupo muy numeroso de vocablos que resultaron del experimento de este trabajo está constituido por adverbios. Se trata de una de las categorías de tipo variable, es decir, es clase abierta y, por lo tanto, está constituida por vocablos menos gramaticales. Pero las formas de la clase verdaderamente abierta se construyen sufijándoles *~mente*, lo que no es verdad de la mayoría de los adverbios del apéndice. En la tabla 4.15 se exhiben todos los adverbios que allí se encuentran, agrupados en los tipos en que normalmente se clasifican. Al final de cada

Tabla 4.15: Tipos de adverbios

tiempo	espacio	modo	cantidad	afirmación	negación	duda
ya	ái, ahi [sic]	así	más	sí	no	qué
aún		mal	muy	siempre	nada	cómo
hoy	ahí	bien	tan	también	jamás	dónde
ora	acá	como	algo		namás	por qué
ayer	allá	cómo	casi		nomás	quizá
ahora	allí	mejor	nada		nunca	
antes	aquí	igual	poco		tampoco	
jamás	onde	claramente	sólo			
luego	abajo	exactamente	menos			
nunca	donde	directamente	mucho			
orita	dónde	precisamente	namás			
tarde	fuera	completamente	nomás			
cuando	junto	perfectamente	tanto			
mañana	dentro	relativamente	apenas			
ahorita		realmente	bastante			
después			siquiera			
siempre			demasiado			
todavía			solamente			
entonces			totalmente			
actualmente			únicamente			

serie están los sufijos en *~mente* que, al considerar que constituyen una clase abierta, son pocos. De estos, casi todos son de modo, aunque hay tres de cantidad y uno temporal. La mayoría de los demás ni siquiera son analizables. Si acaso, están las formas en diminutivo, 'ahor-ita' y 'or-ita', formas con prefijos, 'a-bajo' y 'a-penas', y algunos compuestos, por ej. 'toda-vía', 'tam-poco', 'tam-bién', 'na-más' y 'si-quiera'. Aunque no están todos (por ej. el paradigma de los interrogativos está incompleto), podemos decir que son los miembros de una clase cerrada dentro de una abierta.

Conviene destacar que algunos de estos adverbios pertenecen a más de un tipo, según los contextos en los que ocurran. La tabla 4.16 muestra cuáles son los que se repiten. Lo interesante es que no importa que estén fuera de sus contextos para identificar si pertenecen a más de un tipo (de hecho, según sus contextos pueden adquirir otros significados). Esto no es

Tabla 4.16: Rasgos de algunos adverbios

adv.	tiempo	espacio	modo	cantidad	afirmación	negación	duda
jamás	+				+	+	
nunca	+					+	
siempre	+				+		
dónde		+					+
cómo			+				+
nada				+		+	
namás				+		+	
nomás	+		+	+		+	

solamente porque dentro de sus significados esté especificado que pertenecen a varios grupos[72]. sino que parecen haberse desgastado semánticamente lo suficiente como para aplicarse en diferentes contextos sin causar extrañeza en quien escucha o lee. Por ejemplo, no es sencillo determinar la diferencia entre los adverbios de cantidad 'nomás' (no + más) y 'namás' (nada + más). Pero tampoco son sinónimos. El primer adverbio parece ser más polisémico que el segundo. Por ejemplo, 'nomás acabo' significa 'tan pronto como acabe' (significado temporal) y 'nomás mirando' se refiere a la manera de mirar (de modo). Aunque bien podría sustituirse 'nomás' por 'namás' en estos ejemplos, sin modificar sus sentidos, parece que 'namás' queda mejor en contextos de cantidad, porque el sentido de 'nada' (cantidad nula) está muy presente. Si esto es cierto, 'nomás' es más polisémico, ocurre en más contextos y, por lo tanto, está más desgastado semánticamente. Esto lo haría más gramatical (aunque 'namás' haya sufrido mayor desgaste fonológico). De hecho, esto se refleja en los rangos de la tabla del apéndice: 'nomás' tiene menor rango (152) con un índice de 115.4 varrones, mientras que 'namás' tiene uno mayor (412) con sólo 48.08 varrones de pegamento.

Otra manera de organizar los adverbios para determinar cuántos rincones de lo gramatical cubren, es la identificación de oposiciones entre éstos. En la tabla 4.17 se exhiben las que

[72]Como es el caso de los adverbios de negación que pueden pertenecer a otros tipos (por. ej. 'nada' es también de cantidad.

encontré entre los adverbios identificados como más gramaticales. De nuevo, los huecos se

Tabla 4.17: Oposiciones entre adverbios

positivo/negativo	espacio-temporales	oposición triple
sí - no	antes - después, luego	ayer - hoy - mañana
siempre - nunca, jamás	entonces - ahora/ora/ahorita/orita	aquí - ahí/ái - allí
también - tampoco	tarde - [temprano, pronto]	sí - quizá - no
bien - mal	abajo - [arriba]	antes - ahora - después
mejor - [peor]	dentro[, adentro] - fuera[, afuera]	
más - menos	acá - allá	
más - namás, nomás	[adelante, delante - atrás, detrás]	
mucho - poco	[cerca - lejos]	

marcan entre corchetes cuadrados. No sé cuantas oposiciones no se contemplan aquí, pero de esta tabla algunas no están completas y faltan dos ('cerca' *versus* 'lejos' y 'adelante' y 'delante' *versus* 'atrás' y 'detrás').

Lo importante del grupo de adverbios aislados en este trabajo es, por un lado, que no son los típicos de clase abierta que se construyen con el sufijo *~mente* y, por el otro, que constituyen un grupo que abarca ciertos rasgos modales, cuantitativos y espacio-temporales tan pertinentes a la realidad del hablante, que en otras lenguas suelen codificarse gramaticalmente.

Después del promedio de rangos de los adverbios, el promedio de los de las formas verbales es el más alto. En la tabla 4.18 se agrupan por verbo todas las formas flexionadas que aparecen dentro de las 500 más gramaticales. Los verbos están ordenados por número de formas atestiguadas en la lista. Como puede verse, al final están agrupados algunos de apariencia menos gramatical. Faltan allí cinco formas ('dejan', 'buscando', 'baile', 'tomar' y 'resulta') con rangos más bien altos (alrededor de 400) que no se acomodan en ninguno de los grupos[73]. De hecho, por lo menos los últimos tres grupos de la tabla 4.18 ya constituyen

[73]Se puede argüir que 'baile' es también de movimiento, pero nótese que ese grupo consiste en verbos que

Tabla 4.18: Los verbos y sus formas flexionadas

v.	c.[a]	formas flexionadas
ser	17	fue. es. son. sea. fueron. era. soy. eran. eres. sean. será. serán. sería. somos. siendo. ser. sido
hacer	15	hace. hago. hacen. hizo. haga. hice. hacía. hacían. hacemos. hacerlo. hacerse. hicieron. haciendo. hacer. hecho
haber	12	ha. han. hubo. haya. he. has. había. hemos. habían. hubiera. haber. hay
estar	11	está. están. estaba estoy. esté. estás. estuvo. estaban. estamos. estar. estado
poder	11	pudo. podrá. puedo. puede. podía. pueda. podría. pueden. puedes. podemos. poder
tener	10	tenga. tenía. tienen. tenían. tengo. tiene. tienes. tenemos. tener. tenido
ir	6	fue. fuera. fueron. ir. va. van
querer	6	quería. quiero. quisiera. quieres. quiere. quieren
dar	5	da. dan. dio. dando. dar
deber	4	debe. deberá. deben. debemos
quedar	3	queda. quedan. quedó
seguir	3	sigue. seguir. seguido
de movimiento	6	ir. fue. fuera. fueron. va. van
	3	anda. andaba. andan
	3	sale. salió. salir
	4	pasar. pasaron. pasó. pasado
	2	viene. vienen
	2	baja. bajo
	3.33	
experiencia sensible/mental	3	siente. sentir. sentido
	2	saben. saber
	2	viendo. visto
	1	conocido
	2	
lenguaje	*1.33*	hablar: dicha. dicho; escrito
sistema económico	*1.16*	comprar: cuesta; empleado; ganado; trabajando. trabajo: vale
ciclo vital	*1.16*	vivo. vivir: existen; comer; necesita; cuidado: muerto

[a]Cantidad de formas flexionadas por verbo. Los grupos de más de un verbo presentan el promedio en bastardillas.

una especie de residuo.

No es necesario promediar los valores de glutinosidad total de las formas para corroborar que los verbos que esperaríamos como más usados gramaticalmente se exhiben al principio de la tabla. No se pueden hacer juicios tajantes (de hecho. creo que 'haber' debería estar un poco más arriba, pero no sé si antes o después de 'ser'). De todos modos, hay a grandes

implican desplazamiento. Además. es más prominente como forma sustantiva.

rasgos una progresión de lo más a lo menos usado gramaticalmente.

Algo que no está destacado en la tabla 4.18 son las formas verbales para expresar la existencia o presencia de las cosas. Esta es una función lógica que esperaríamos en todas las lenguas. Aquí se encuentran, además de las formas de 'ser', 'estar' y 'haber', las representativas de este último ('hay' y 'hubo') y una de las formas de 'existir'. Además, es curioso que la única con enclítico pronominal es 'hacerlo' que, sin existir la categoría *proverbo*, hace las veces de verbo pleno con el significado de la acción verbal previamente referida en el contexto.

Con respecto a los adjetivos y sustantivos que ocurrieron en la lista de los vocablos más gramaticales, los examinaremos como un solo grupo, especialmente porque, aunque haya una diferencia en promedios de rangos (véase la tabla 4.11), los primeros pueden fungir pronominalmente como los segundos. En la tabla 4.12 se presentaron los adjetivos determinativos que ocurrieron en esa lista y que típicamente ocurren antes del nombre o como pronombres. Los adjetivos que se examinarán junto a los sustantivos son los calificativos, típicamente pospuestos al sustantivo.

En la tabla del apéndice se observa una variedad importante de vocablos sustantivos y adjetivos. Dada su heterogeneidad podemos presumir que se trata de una muestra aleatoria de vocablos más o menos frecuentes (el menos frecuente es 'oscuro', núm. 468 con 70 ocurrencias). Sin embargo, hay varios grupos relacionados con ciertos aspectos de la realidad muy a menudo codificados gramaticalmente. Lo interesante es que algunos de estos constituyen conjuntos ordenados que representan conceptos muy generales, es decir, representan una especie de paradigmas.

Por ejemplo, los conceptos de tiempo y espacio son contenidos típicos de preposiciones y

de los adverbios más gramaticales (véase la importancia de estos rasgos en las tablas 4.15, 4.16 y 4.17). Ni hablar del importante papel del parámetro temporal en la flexión verbal del español. Por eso no es tan casual que entre los resultados haya varios sustantivos y adjetivos que codifican este tipo de información. En la tabla 4.19 se muestran las formas

Tabla 4.19: Nociones espacio-temporales

tiempo	vez veces tiempo mañana tarde horas día días años inicial anterior actual presente pasado viejo joven nueva nuevo nuevas nuevos
espacio	inferior superior anterior media central casa sociedad público personal
series	primer primera primeras primeros segunda segundo tercera inicial próximo siguientes último última
tamaño	mayor mayores gran grande grandes alta alto menor pequeño pequeñas pequeños baja bajo
cantidad	solo sola única varias varios total completa
posesión	propia propio

que codifican cuestiones relacionadas con el tiempo y el espacio. Hay algunas que suelen servir para codificar ambos rasgos (por ej., 'anterior'), aunque en general esas lecturas se desprenden de los contextos en que ocurren. Además, hay nociones como el orden de las cosas que organizan objetos o hechos en series espaciales o temporales.

El tamaño también está relacionado con la noción de espacio. En la misma tabla se agrupan las formas con algún contenido relacionado con este concepto. Recuérdese que en México algunos de estos adjetivos también se asocian con la edad, es decir, con el tiempo (**'es una persona grande'** se refiere a menudo a la edad de la persona en cuestión).

La otra noción espacio-temporal importante es la de cantidad, que ya documentamos como de uso gramatical importante al ocurrir como concepto organizador de los adverbios. En la tabla 4.19 se listan las formas adjetivas que cuantifican objetos (por ej. 'única', 'varias',

etc.) y sus partes o proporciones ('total' y 'completa').

Finalmente está la posesión. Recuérdese que en español existen paradigmas de adjetivos determinativos y pronombres que indican posesión (tablas 4.12 y 4.13), cosa que hace patente lo prominente de este concepto en esta lengua. Por eso, no sorprende que por lo menos el lexema 'propio' esté representado entre las formas en alguna medida gramaticales.

Aparte de las formas con contenido espacio-temporal, es natural que al tener sustantivos en la muestra, encontremos formas para designar objetos. La tabla 4.20 muestra una posible y esquemática configuración de los objetos representados. Entre las cosas que se encuentran

Tabla 4.20: Objetos concretos y abstractos

entidad	cosa cosas forma elementos grupos partes sistema
cuerpo	ojos corazón
información	datos información ideas dicho escrito
varios	carácter condiciones problemas

se pueden distinguir nombres de 'cosas' muy abstractas, aptas de referirse a cualquier entidad. De hecho, casi podrían considerarse anáforas para referirse a los asuntos que se repiten mucho en el discurso. Además, el hecho de que dos formas que representan partes del cuerpo estén presentes en esta tabla es interesante porque se ha observado que suelen convertirse en signos gramaticales de algún tipo (por ej. preposiciones).

Por otra parte, las formas que representan procesos son en realidad muy pocas, 'actividad', 'trabajo', 'baile' y 'estudios', pero constituyen una muestra de términos en relación de superordinación (sobre todo los dos primeros) con muchas otras formas que representan procesos específicos[74].

[74] Las formas 'estudios' y 'baile' parecen muy específicas, pero se refieren, según los entornos en que ocurren, a una gran variedad de tipos actividades: hay muchas cosas que se estudian y no sólo bailan las personas.

Así como la presencia de sustantivos implica la existencia de objetos entre las cosas representadas por las formas de la tabla del apéndice, la presencia de adjetivos es motivo para suponer la aparición de propiedades para definirlos. En la tabla 4.21 se organizan las propiedades que se encontraron representadas entre las 500 formas del apéndice. De nuevo, encontramos formas con significados muy abstractos o muy representativos de las preocupaciones más comunes de los hablantes. En el primer grupo hay formas adjetivas

Tabla 4.21: Propiedades

identidad	cierta ciertas ciertos diversas diversos distintas distintos igual diferentes
física	abierto, frío
tiempo	joven viejo nueva nuevo nuevas nuevos (etc.)
ontología	libre posible
color	blanca negra oscuro verde
experiencia mental	feliz artística loco conocido sentido
calificativos	bueno buena buenas fácil especial mala normal común general constante principal fundamentales necesario

importantísimas para determinar objetos de cualquier naturaleza. Las formas adjetivas 'igual' y 'diferentes' aluden a la función lógica para establecer la identidad o no identidad entre las cosas. Otro par de formas interesante es el que se refiere a las propiedades de libertad y posibilidad que en español están gramaticalmente codificadas en el verbo 'poder'.

Aunque pobremente, el color y algunas experiencias mentales están también representados en la tabla 4.21 (no se consigna allí el sustantivo 'dicha' que en una de sus acepciones es sinónimo de felicidad y sí aparece en la lista de las 500 formas). Al considerar la variedad de colores y vida mental de los hablantes de una cultura, asuntos tan presentes en el discurso, que tal vez pudieran abrirse paso hasta la codificación gramatical[75], esta muestra es muy

[75]Recuérdese que en la tabla 4.18 se consignaron 8 formas verbales relacionadas con la experiencia mental, aunque, en realidad, sólo esas formas tampoco revela mucho.

escasa. Son más interesantes las últimas formas de la tabla (bajo el rubro "calificativos"). Se trata de adjetivos abstractos para juzgar si algo es bueno, común o necesario, cuestiones que obviamente preocupan a los hablantes de muchas lenguas.

Es obvio que mientras más cosas se quieran incluir entre lo gramatical, menos gramatical va a ser lo que se incluya. De hecho, mezclados con los vocablos gramaticales encontramos hacia el final de la tabla del apéndice una variedad de signos cuyo estudio merece más un acercamiento semiótico. Se trata de cosas que caracterizan las vivencias más generales de los mexicanos, mucho más que de signos que sirvan para organizar y matizar el discurso. En este sentido, tenemos una colección de formas léxicas con contenidos obviamente culturales. Si hasta aquí se han diversificado considerablemente las posibilidades de agruparlas, lo que

Tabla 4.22: Formas léxicas

naturaleza	agua tierra natural puro pura vida muerto
mundo	país países internacional mundo
sociedad	cultura cultural popular moderna mexicana mexicano mexicanos nacional nacionales partido política político público estado social sociedad trabajo problemas militar
economía	comercial comerciales dinero pesos económica económico actividad empleado ganado industrial mercado pobre problemas productos tierra trabajo profesional valor
personas	gente pueblo personas humano humana hombre mujer mujeres madre papá hijos mexicano mexicana mexicanos empleado personal pobre viejo joven loco negra

queda se puede ordenar de muchas más maneras, muy sujetas a interpretaciones personales[76], como en la tabla 4.22. Como se ve, ya no son signos gramaticales. Si estas formas tienen alguna función, es la de referirse a las cosas de la realidad.

De todas maneras, son lo suficientemente abstractas como para que sus significados varíen

[76]Por ejemplo, se puede argüir que estas formas representan un conjunto estructurado de signos muy generales pero imprescindibles para entender la cultura mexicana, pero este tipo de interpretaciones son más bien forzadas, principalmente porque los signos están fuera de sus contextos.

considerablemente de contexto en contexto. Además, varios mantienen numerosas relaciones de superordinación con lexemas que no están allí: hay muchos tipos de 'productos', de 'actividades', 'personas', 'mexicanos'[77], 'problemas', etc. Sería interesante determinar los lexemas más hiperonímicos del corpus y contrastarlos con los de las de esta tabla.

Por último, cabe especular acerca del futuro. Si nuevas formas gramaticales han de emerger en el porvenir lejano, estos segmentos léxicos son buenos candidatos para convertirse —después de sufrir los desgastes semánticos y fonológicos que el destino les depare— en signos función de algún tipo. Por ejemplo, de entre ellas puede emerger (ya que se trata de sustantivos y adjetivos) un conjunto de clasificadores nominales que refleje los tipos de cosas que les será importante a los mexicanos del futuro.

Hasta aquí, si no se me escapa nada, hemos examinado todos los vocablos de la tabla del apéndice. Es evidente que todavía falta investigar estos signos gramaticales, aislados mediante un procedimiento de selección cuantitativa aplicado a un corpus. Como vimos, las formas gráficas de la tabla D.1 constituyen un grupo heterogéneo de vocablos con referentes muy diversos. Pero no son un conjunto aleatorio de palabras. Hay cierta organización implícita en su presencia. Por eso se pueden reagrupar alrededor de ciertos rasgos de la realidad frecuentemente codificados gramaticalmente. Esto permite establecer su carácter de conjunto organizado.

También falta analizar lo que aparentemente no debería estar allí. Hay muchas cosas que no entran dentro de los catálogos de palabras funcionales de las gramáticas españolas, pero

[77]Especialmente en un corpus del español de México, múltiples referencias a una gran variedad de tipos de 'mexicanos' son de esperarse.

que, por su variedad y restricciones de uso, tampoco parecen palabras de contenido típicas. Si bien muchos segmentos se reconocen inmediatamente como elementos funcionales, otros, aunque de entrada no lo sean, requieren de la exploración de los contextos en donde ocurren para apreciar la medida en que su carácter es o no gramatical.

Sin embargo, el que una porción importante de los vocablos gráficos de la tabla del apéndice formen parte de los paradigmas fundamentales de palabras gramaticales bien conocidas y estudiadas de la lengua española (algunas de las cuales exhiben frecuencias bajísimas) indica que su uso gramatical puede medirse mediante esquemas más elaborados que la simple determinación de frecuencias.

4.6 Observaciones finales

En este último apartado se resumen los conceptos presentados en este capítulo. El capítulo se inició con una breve presentación de algunas intuiciones de diversos lingüistas que apuntan hacia la existencia de una fuerza o energía que, muy aparte de sus posibles nombres y adjetivos, existe entre las palabras de una lengua y que se manifiesta en la diversidad de los objetos lingüísticos a través del tiempo. En segundo lugar, se examinaron algunas cuestiones generales sobre la relación entre afijalidad y cliticidad para concebirlas como cualidades similares de objetos lingüísticos de diferentes niveles (al interior y al exterior de las palabras) y, por lo tanto, como instancias de una fuerza más general de adhesión entre fragmentos de palabras, entre palabras y entre palabras y fragmentos.

Aunque en este trabajo se ha considerado esta fuerza solamente en su dimensión

sincrónica, lo que significa que cada objeto lingüístico se caracteriza mediante dos valores únicos de glutinosidad (uno de su derecha y otro de su izquierda), dicha fuerza también debe manifestarse en la dinámica del discurso y, no menos importante, en estados de lengua distintos.

En tercer término, se presentaron los conceptos principales de la teoría de medición, con el objeto de determinar las características necesarias de un esquema que nos permita rendir cuenta del fenómeno de glutinosidad. Además de contemplarse los conceptos de dimensión, escala, unidad, errores, etc., se enunciaron los axiomas en los que debe basarse una glutinometría formal y coherente. En concreto, se estableció que cuantificar una propiedad consiste en hacer corresponder los grados de una propiedad g^o al conjunto de números r de tal manera que el orden y espaciamiento de los números refleje el orden y espaciamiento de los grados. Entonces, el acto de medir consiste en determinar algunos de esos valores numéricos; las magnitudes son representaciones numéricas precisas de los grados de una propiedad. La medición es la parte empírica de la cuantificación y consiste en interpretar las marcas en un instrumento como números que proporcionan una imagen más o menos cercana a los grados de la propiedad. Todo se resume en determinar empíricamente cuántas unidades o subunidades caben en una magnitud. Esto es, la medición propia requiere —como no lo requiere la cuantificación— de la selección de unidades que midan la propiedad y de una escala compleja (tanto material como conceptual) que represente a las unidades en relación a un origen.

De esta manera, en este capítulo se propusieron unidades para medir las magnitudes involucradas en el cálculo de la glutinosidad y se estableció al *Varrón* como unidad de la

glutinosidad, es decir. para medir la cantidad de información que fluye al interior de un signo formado por signos del nivel inferior en relación económica. En la tabla 4.7 se resumen las unidades y dimensiones involucradas en la medición de la glutinosidad.

Finalmente, en las últimas secciones se examinaron los resultados de la aplicación de todos estos conceptos en el *Corpus del Español Mexicano Contemporáneo*, se presentaron los signos gramaticales más importantes y, con el objeto de corroborar su carácter de uso gramatical. se clasificaron en grupos para exhibir los rincones que ocupan en la gramática española.

morfología automáticas que sirve de marco para el desarrollo de los capítulos siguientes. Así, después de examinar brevemente las bases de los formalismos gramaticales más conocidos y examinar los métodos de adquisición léxica más conocidos, se revisaron algunos enfoques para estudiar la morfología mediante computadoras, especialmente los métodos más conocidos de segmentación automática de morfemas que se desarrollaron a partir de la mitad del siglo XX.

En el segundo capítulo se presenta la investigación del nivel morfológico, destinada a llevar a cabo los primeros dos objetivos de la tesis, es decir, determinar automáticamente un conjunto de signos al interior de la palabra y construir un programa segmentador de vocablos a partir de ese conjunto. Para eso, se presenta una discusión de procedimientos de descubrimiento de unidades morfológicas, y se determina al afijo como el tipo de unidad morfológica más apta de ser investigada automáticamente. Después se hace una investigación de los afijos en el *CEMC*, con el objeto de falsificar la hipótesis sobre afijalidad de segmentos de palabras propuesta en la introducción.

El tercer capítulo —"El clítico en el *CEMC*"— se ocupa de la investigación cuantitativa al exterior de la palabra gráfica para descubrir las palabras más gramaticales del español mediante el *CEMC* (objetivos tercero y cuarto de la tesis). En esencia se aplican los métodos explorados en el capítulo anterior para descubrir afijos. También se examina el fenómeno de la puntuación como indicio de fronteras sintagmáticas. Luego, la investigación empírica se centra en el descubrimiento de clíticos y otros tipos de signos gramaticales a partir del *CEMC*: se busca verificar o falsificar la hipótesis sobre la cliticidad entre palabras gráficas.

En el último capítulo se investiga un esquema generalizador de los métodos aplicados en los primeros capítulos. Primero se exploró la relación entre afijalidad y lo que en un principio

se llamó cliticidad para proponer un índice cuantitativo capaz de medir la energía o fuerza de enlace que une a cada objeto lingüístico con sus contextos en un corpus y que aquí fue llamada glutinosidad. Después se examinan los conceptos y requisitos generales de la teoría de la medición para construir un esquema glutinométrico probablemente aplicable no sólo al español, sino también a otras lenguas, cuando menos semejantes en sus propiedades de entropía y economía.

La parte experimental se llevó a cabo sobre todo en los capítulos segundo y tercero. En el último sólo se repitieron los experimentos más importantes con el objeto de reflejar las intuiciones que allí se elaboraron para construir una glutinometría. Aunque brevemente, a continuación se describen con más detalle los experimentos de esta investigación.

4.6.1 Sinopsis de experimentos

En el segundo capítulo se consignan los experimentos morfológicos. Estos consistieron principalmente en explorar y comparar varios métodos para segmentar vocablos. En concreto, se aplicaron las estadísticas de no asociación para digramas (prueba de χ^2, información mutua, razón de semejanza, y el coeficiente de Yule); se contaron para cada segmentación las estructuras combinatorias llamadas cuadros que se atestiguaron en el corpus: se midió la cantidad de información inherente a los segmentos que forman los vocablos; y se estimó el nivel de ahorro al sistema según las relaciones económicas de dichos segmentos. Mediante una muestra de alrededor del 1% del número total de vocablos en el corpus (cerca de 79.000), se corroboró la intuición de que los últimos tres métodos resultaban en mejores segmentaciones porque capturan la esencia de lo que concebimos como afijo.

Conclusiones

> La estadística, ya digo, ha suplantado a la interpretación, y por tanto ha desterrado los matices, las sutilezas, las complejidades de todo lo humano, hasta la sensatez ha suplantado. Nadie se extraña ya, ni se ríe, cuando lee idioteces y absurdos del tipo: "Los españoles tienen 1.34 hijos por pareja", como si los hijos pudieran en verdad ser troceables y fuera posible que alguien tuviera efectivamente 1.34 vástagos.
> Parece haberse olvidado que para elaborar una de esas sacrosantas estadísticas es preciso, por principio, reducir y simplificar la realidad al máximo, y por tanto falsearla; de modo que la estadística sería, en el mejor de los casos, algo vagamente orientativo y —en contra de lo que se cree— obligadamente inexacto.
> **Javier Marías**

En esta última sección se resumen los resultados de esta investigación y se exponen las conclusiones. Después de un breve recuento de la tesis, se presenta una sinopsis de los experimentos llevados a cabo en los dos primeros capítulos. Luego se enumeran las desventajas del enfoque propuesto, así como los problemas encontrados y las cosas que quedaron pendientes a lo largo de este trabajo. Después se analizan las ventajas y se listan los logros del enfoque, en especial con respecto a lo propuesto en la introducción: los objetivos iniciales y las hipótesis de investigación —de afijalidad, cliticidad y glutinosidad— que sirvieron de espina dorsal de la tesis, todo esto con el objeto de reformularlas para investigaciones futuras. Finalmente, se enumeran en el último apartado las conclusiones principales del trabajo.

Además de estas observaciones finales y una sección introductoria, la tesis consiste de cuatro capítulos. El primero es un panorama general de los métodos de la sintaxis y la

A partir de estos resultados, se procedió a construir un catálogo[78] de afijos del español y un programa para segmentar palabras en bases y afijos. Aquí fue donde se sometió a prueba la hipótesis de afijalidad al calcular para cada segmento el índice normalizado que refleja dicha hipótesis. El hecho de que los valores más altos fueran consistentemente asignados a segmentos que representan a afijos conocidos y a cadenas frecuentes de afijos del español no permitió que la hipótesis se falsificara.

En el tercer capítulo se describen los experimentos para examinar la hipótesis de cliticidad, la cual se simplificó hasta apoyarse solamente en los valores de entropía y economía. De esta manera, para cada palabra gráfica se calcularon dos índices normalizados, uno para cada uno de sus lados. El hecho de tener dos valores para cada segmento permitió ordenarlos de distintas maneras, sobre todo para explorar la mecánica de cómo se pueden relacionar estos valores en el examen de la naturaleza gramatical de cada segmento. Aquí quedó claro que la hipótesis de cliticidad no es exactamente eso, porque consigna una medida no tanto de qué tan clítica es una forma, sino de su importancia en una escala de asociación de tipo más general (digamos que de asociación gramatical). Así, por un lado, la propiedad de ser un clítico es una función de los dos valores de cada segmento (mientras mayor es su diferencia, mayor es su cliticidad). Por el otro, la propiedad de formar parte del conjunto de formas que, sin ser clíticos, le dan estructura al discurso (las más gramaticales) es también una función de ambos valores (mientras mayor es su suma, más obvio es su carácter de palabra función de la lengua del corpus). De esta manera, aunque la relación entre entropía y economía establecida en la hipótesis se sostiene como una medida pertinente de algo lingüístico, no podemos llamarlo

[78] En realidad se construyeron varios catálogos de afijos, según diferentes modificaciones de caracteres (véase el apéndice sobre el *CEMC*).

cliticidad. En ese sentido se falsificó dicha hipótesis, cosa que necesariamente se refleja en la reformulación presentada más abajo.

Otro experimento importante de este capítulo fue el cálculo de un índice de puntuación que incide en el fenómeno de asociación entre palabras gráficas. Es decir, se tomó en cuenta la aparición de caracteres no alfabéticos para calcular una medida de lo que se interpone entre segmentos no alfabéticos, la cual debe naturalmente ser baja entre un clítico y las palabras a las que se adhiere. Este índice dio tan buenos resultados, que amerita considerarse, si bien no dentro de la dinámica general de la asociación lingüística, sí dentro de un fenómeno cuantificable de la lengua escrita, cuando menos útil en el descubrimiento de fronteras sintagmáticas.

En cuanto a la relación entre afijalidad y lo que llamamos en un inicio cliticidad, espero haber aclarado que, a pesar de la diferencia conceptual, la primera (que mide los puntos de menor asociación lingüística entre los segmentos de las palabras gráficas) y la segunda (que mide las uniones entre palabras gráficas de mayor asociación), los dos índices son instancias de lo mismo, es decir, los dos son directamente proporcionales y miden una especie de energía (como la llamó Sapir) de adhesión o pegajosidad entre segmentos de un corpus, que podemos llamar glutinosidad.

En suma, la afijalidad de un segmento se puede concebir como una combinación de ciertas dimensiones medibles entre dos segmentos de palabra. La cliticidad, por otra parte, también se puede describir en razón de estas dimensiones, pero al exterior de lo que hemos definido como palabra y mediante la comparación de los valores obtenidos a cada lado del segmento examinado. Por un lado, para que un segmento sea afijal, se espera una segmentación

económica y una alta entropía (ya sea en una u otra dirección, cosa que define el tipo de afijo). Por el otro, el índice de cliticidad de las palabras se calculó esperando también una alta entropía y un índice alto de economía. Pero, como se vio, un clítico no se determina por el valor aislado de este índice, sino mediante la comparación de índices de este tipo para cada lado del segmento. Similarmente, al considerar que la afijalidad y la cliticidad son paralelas, cabe suponer que la primera también es una función de dos índices: el índice de afijalidad al interior de la palabra y el índice de glutinosidad del segmento afijo con respecto al exterior de la palabra.

Dadas todas estas observaciones, las hipótesis de este trabajo se pueden reformular para reflejar lo observado en estos experimentos. De hecho, se puede resumir todo en una sola hipótesis de glutinosidad que abarque los fenómenos asociación gramatical tanto al interior como al exterior de la palabra gráfica, de tal manera que los fenómenos de afijalidad y cliticidad podrían considerarse sus consecuencias.

4.6.2 Las hipótesis de glutinosidad reformuladas

Hay una fuerza de enlace o glutinosidad entre los segmentos más gramaticales y los segmentos léxicos de un corpus que es directamente proporcional a la entropía que disparan los primeros y al número de signos del nivel siguiente producidos mediante la combinación de los gramaticales con los léxicos. Concretamente, la glutinosidad medible a cada lado de un segmento s_x en el corpus es directamente proporcional al producto de la incertidumbre que provoca en esa dirección (h_x) por el ahorro o economía que le significa al sistema (k_x) dada la misma dirección: $GL(s_x) = h_x k_x$. Para ser más exactos, a la derecha de s_x hay una

glutinosidad, $GL^d(s_x)$, estimable mediante los valores de entropía y economía de izquierda a derecha a partir de s_x (esto es, $GL^d(s_x) = h_x^d k_x^d$); y a la izquierda otra, $GL^i(s_x)$, estimable mediante los valores de entropía y economía de derecha a izquierda a partir de s_x (esto es, $GL^i(s_x) = h_x^i k_x^i$).

Además, si cada segmento tiene dos medidas de glutinosidad (una a la derecha y otra a la izquierda), su carácter estructural en el corpus (es decir, su propiedad de fungir como elemento gramatical) es directamente proporcional a la suma de estas medidas. Por otra parte, tanto la afijalidad como la cliticidad de cada segmento son directamente proporcionales al exceso de glutinosidad que haya de un lado con respecto al otro. La primera propiedad exhibe simplemente valores mayores que la segunda.

Todo esto apunta a que el carácter gramatical de los segmentos de un corpus es una función de sus relaciones con el resto de los segmentos del mismo y que puede, por lo pronto, entenderse como un flujo de información (oscilación entre certidumbre e incertidumbre, entre lo familiar y lo inesperado) suscitado y constreñido por una estructura económica de signos.

4.6.3 Problemas, ventajas y pendientes de una glutinometría

Primero, la gran desventaja, no nada más de este enfoque, sino del trabajo con córpora en general es que su recolección es muy cara. El problema es aún mayor si los córpora recolectados son para procesamiento computacional. A pesar de las nuevas tecnologías que seguramente en un futuro solucionarán mucho de esta carga, construir un corpus sigue requiriendo mucho tiempo y dinero, por no hablar del gran esfuerzo humano. Sin embargo, los lingüistas (algunos más, otros menos) de todos modos tenemos que recolectar información de

diversas fuentes para compilar nuestros córpora, por pequeños que sean (hay quien hasta se los inventa). En ese sentido, la verdadera desventaja es que la intuición y la introspección solas no nos permitan conocer todo lo que quisiéramos sobre el lenguaje.

En cuanto a establecer técnicas para medir fenómenos lingüísticos en córpora, el carácter especulativo de la lingüística es otra desventaja. De hecho, entre las teorías especulativas, como las sociales, los intentos de establecer técnicas de medición de carácter estándar han tenido poco éxito fuera de los estudios económicos y demográficos. Algunos de los problemas son la ausencia de marcos teóricos aceptados universalmente y, por lo tanto, de mesurandos cuantificables (contables), errores de muestreo y problemas asociados a la intrusión de quien mide en cuanto al objeto social que se mide y el carácter subjetivo de la información proporcionada por los seres humanos. Pero como el lenguaje también tiene un carácter esencialmente matemático (digamos que la glutinosidad no es un fenómeno social), hay mucho lugar en la lingüística (mucho más que, por ejemplo, en la psicología) para utilizar técnicas que midan los fenómenos lingüísticos. Todo está en decidir qué y cómo se quiere medir. Los problemas de representatividad, muestreo y la intrusión del investigador (quien, por ejemplo, levanta datos en un comunidad) son problemas que de todos modos se presentan en todas las disciplinas, por poco especulativas que sean.

Por otra parte, un enfoque de este tipo también tiene sus ventajas. Primero que nada, al concebirse y cuantificarse el fenómeno de glutinosidad, este esquema tiene cuando menos las ventajas de la cuantificación en general que, como apunta Bunge, son: refinamiento conceptual, descripción más precisa y clasificación más precisa[79]. Por ejemplo, el concepto

[79] Bunge, *op. cit.* [25] 1967, p. 202.

de *gramaticalidad* queda definido de manera muy precisa, no como un juicio absoluto impuesto a un fenómeno relativo, ni como una intuición educada sobre la fosilización de un segmento, sino como un valor numérico abstracto, pero muy preciso, que caracteriza a cada segmento de un corpus y que se calcula de manera concisa a partir de otros valores específicos. Sobra decir que este refinamiento conceptual sirve para describir y clasificar los objetos de un corpus de manera muy precisa.

Esto resulta en otras virtudes del enfoque. La más inmediata es la posibilidad de descubrir automáticamente las unidades lingüísticamente más pertinentes de un corpus sin conocer la lengua allí representada. Claro que esto no excluye otras técnicas, sobre todo las cualitativas, pero constituye un buen marco de dónde partir. Otra ventaja es la posibilidad de conocer de otra manera las unidades lingüísticas de lenguas muy estudiadas. Aunque sean muy conocidas, al observarlas y clasificarlas mediante representaciones numéricas abstractas que las caractericen como miembros de un sistema, estas unidades se pueden investigar y describir de manera cuantitativa. De hecho, sus representaciones numéricas constituyen juntas una descripción cuantitativa del sistema al que pertenecen.

Además, este tipo de investigación abre muchas puertas para el estudio del lenguaje al permitir que se puedan medir otras propiedades lingüísticas. Por ejemplo, al comparar estados distintos de una misma lengua se puede obtener evidencia cuantitativa dura (más allá de intuiciones y corazonadas) de los procesos de gramaticalización del tipo concebido por Meillet.

Pero muy aparte de las ventajas y desventajas del enfoque, hay todavía muchos asuntos que resolver. Por ejemplo, en cuanto a los afijos, valdría la pena llevar a cabo las siguientes

tareas:

1. Estudiar las diferencias entre prefijos y sufijos, sobre todo en lenguas que, en su morfosintaxis, dependan más de los primeros. Si en español los primeros parecen estar más integrados a la raíz, sería interesante examinar estas lenguas para ver si los procedimientos para descubrir prefijos funcionan mejor.

2. Investigar la afitáctica. Como vimos en el primer capítulo en lo referente a la morfología de estados finitos, los estudios automáticos ya se han ocupado de la afitáctica de las lenguas, pero sin ocuparse previamente de las unidades morfológicas pertinentes (esto es, las presuponen). Si el aparato morfológico es un sistema, conviene primero determinar sus partes en lugar de creerle a la tradición o adivinarlas. En cuanto a un esquema glutinométrico, valdría la pena investigar cómo funciona la afijalidad con respecto al orden de los afijos.

3. Determinar automáticamente paradigmas. Desde las investigaciones de Andreev, no ha habido intentos por descubrir cuantitativa y exhaustivamente los paradigmas completos de las lenguas. Como tema clave de la lingüística, urge investigarlo automáticamente.

4. Investigar afijos dentro de cada clase de vocablos. Es decir, si el estudio de los clíticos y el conocimiento previo (por inspección del corpus o fiándose en información *a priori*) nos permite clasificar automáticamente los vocablos que le siguen a un grupo de clíticos (por ej., los pronombres proclíticos), se pueden aplicar los procedimientos de segmentación a cada clase por separado bajo la hipótesis de que se afinarían las estimaciones de afijalidad y se resolverían algunas ambigüedades (las de las formas polisémicas que pertenecen a más de una clase).

En cuanto a las palabras gramaticales, entre ellas los clíticos, sería interesante llevar a cabo las siguientes investigaciones:

1. Investigar automáticamente la flexión de las formas más gramaticales que no recurren a formas supletivas. Hay verbos muy gramaticales (por ej., 'haber', 'deber', 'poder', etc.) y otros tipos de vocablos función (por ej. 'cuya', 'cuyo', 'cuyas', etc., incluso determinadores como 'la', 'las', 'esta', 'estos', etc.) que representan paradigmas que también contienen formas supletivas, las cuales dificultan su estudio mediante los procedimientos para las formas regulares.

2. Estudiar la relación entre los pares de glutinosidades de los segmentos de otras lenguas, con el objeto de, por ejemplo, aclarar la naturaleza de los vocablos con valores cercanos entre sí (es decir, con tanta glutinosidad de un lado y del otro).

Por último, para afinar los conceptos de la glutinometría propuesta, quedan pendientes las siguientes cuestiones:

1. Intentar falsificar la hipótesis de glutinosidad (si en efecto se trata de una función de la economía y la entropía de los segmentos del corpus), sobre todo en el marco de los datos de otras lenguas.

2. Continuar investigando las técnicas para medir la glutinosidad. Si llegamos a un acuerdo en cuanto a su existencia y sobre qué fenómenos la determinan, se puede poner énfasis en investigar los mejores métodos para medir dichos fenómenos (es decir, todavía no sabemos si hay mejores que los aplicados en este trabajo).

3. Investigar cómo se concebiría una glutinosidad "no direccional" o "bidireccional", como la que podríamos imaginar al interior de compuestos. Tanto la afijalidad como la cliticidad (y por lo tanto el tipo de glutinosidad que se ha definido aquí) son fenómenos direccionales, es decir, hay una dirección de asociación. Valdría la pena ver cómo podría funcionar y medirse esta fuerza entre las bases de vocablos compuestos.

4. Diseñar mejores herramientas para medir la glutinosidad. Mejores métodos para medirla implican mejores instrumentos[80]. De hecho, los mismos métodos se pueden utilizar en la construcción de instrumentos más eficientes que aprovechen las constantes innovaciones tecnológicas.

5. Estudiar los fenómenos de la puntuación. Si la lengua escrita es un sistema diferente a la lengua hablada, hace falta investigar qué tanto las evidencias de una son pertinentes a la otra. Esto depende indudablemente del tipo de sistema de escritura de la lengua examinada. Aun las lenguas con sistemas de escritura cercanos a los orales y con topogramas especializados en el esclarecimiento de la sintaxis merecen una investigación detallada de la distribución de su subsistema puntuacional antes de tomarlo como algo dado (lo que fue el caso en este trabajo).

6. Investigar métodos para medir errores. La fe y esperanza que se invierte en creer que los valores medibles convergen a los valores reales son el motor para tratar de dilucidar el por qué diferentes técnicas pueden resultar en diferentes estimaciones de lo que una misma teoría puede asumir como un mismo valor y, lo que es más, son la gran motivación para buscar y hacer operativas nuevas técnicas y procedimientos de medición. Como dice Bunge, allí es donde cabe esperar avances científicos significativos[81].

Pero muy aparte de todo lo que quedó pendiente y con respecto a los objetivos planteados

[80] Bunge, *op. cit.* [25] 1967, p. 239.

[81] Bunge, *op. cit.* [25] 1967, p. 209.

en un principio, en este trabajo se lograron varias cosas, que vale la pena enumerar a continuación:

- Se compararon diversos métodos de segmentación de palabras —estadística de digramas (prueba de independencia de χ^2, información mutua, razón de semejanza, coeficiente de Yule), entropía, índices de cuadros y de economía de de Kock— y se mostró que estos últimos (aquellos que miden algún rasgo lingüístico de los afijos en general) son más confiables.

- A partir de estos métodos —que no son específicos para la lengua española—, se propuso y aplicó un índice formal apto para medir la cualidad de afijo (o *afijalidad*) de cualquier segmento de palabra o vocablo.

- Se extendieron estos métodos para medir el carácter de clítico (o *cliticidad*) de aquellos segmentos que, aunque gráficamente independientes (esto es, aparecen entre dos espacios), dependen formalmente de éstas. Se afinó el cálculo de esta propiedad al tomar en cuenta no uno sino dos valores de asociación (de la derecha y de la izquierda): al observar los datos, se hizo evidente que la cliticidad debe ser función de la diferencia de estos valores.

- Se formalizaron estos métodos o criterios. Es decir, se determinaron fórmulas de carácter matemático para extraer los datos morfológicos (relativos a la segmentación morfológica) y morfosintácticos (en relación a la asociación entre palabras y segmentos).

- Basándose en los criterios de *afijalidad*, se construyeron automáticamente diversos tipos de catálogos de signos afijales del *CEMC*.

- De manera similar, se seleccionaron los signos más gramaticales de ese corpus, es decir, aquellos más aptos de funcionar —según los criterios de *cliticidad* examinados— como elementos estructurales por su distribución en dicho corpus. Concretamente, así como la cliticidad verdadera es una función de dos valores de asociación (su resta), el carácter gramatical de cada segmento se puede estimar mediante la suma de dichos valores.

- Los catálogos de afijos, clíticos y las listas de palabras función obtenidas automáticamente a partir del *CEMC* constituyen un tipo de descripción formal de la lengua española, cuando menos de la hablada y escrita en México.

- La investigación de métodos para cuantificar tanto la afijalidad como la cliticidad de los segmentos de un corpus condujo a su generalización en la noción de *glutinosidad* o pegajosidad cuantitativa entre cadenas de morfemas al interior de la palabra y del sintagma. Se propuso un esquema glutinométrico (que especifica unidades de medición, bases axiomáticas, etc.) apto de medir dicha glutinosidad, cuando menos en lenguas como el español.

- Se construyeron diversos programas para los experimentos de descubrimiento y procesamiento de signos gramaticales tanto al interior como al exterior de la palabra gráfica, entre ellos un *tokenizer* o fichador (que filtra y segmenta el corpus, además de contabilizar la puntuación), un separador de raíces y afijos, un programa que ordena los vocablos gráficos según sus índices de entropía y economía y dos glutinómetros (uno para el interior y otro para el exterior de la palabra gráfica).

4.6.4 Conclusiones

El lenguaje es indudablemente infinito, pero esto —como hemos visto— no significa que —al igual que la infinita realidad— no sea susceptible de estudiarse cuantitativamente. En otras palabras, si bien el lenguaje no deja de ser infinito, espero haber mostrado que los métodos cuantitativos sí son herramientas apropiadas para su estudio. Además, hay que enfatizar que los métodos cuantitativos de ninguna manera sustituyen o desplazan al trabajo cualitativo. Muy al contrario, lo apoyan. La cantidad de datos (el tamaño del corpus) es crucial, pero el verdadero problema es el de la representatividad del corpus utilizado.

Mucho se ha hablado de la entropía y la economía como características universales del lenguaje. De todos modos, es impresionante corroborar el hecho del lenguaje como una estructura entrópica y económica, cuyas unidades, por tanto, son susceptibles de investigarse mediante aquellos métodos capaces de medir esas propiedades. Todo esto apunta a que el léxico como sistema tiene una estructura interna que no se debe subestimar al seleccionar un vocabulario estándar. Como dice de Kock, si bien restringir el número de vocablos en un sistema puede en teoría "aliviar" la carga de memorización en los hablantes, lo cierto es que esto puede afectar la economía interna del sistema léxico, incluso en la dimensión diacrónica: "It even more interferes with historical conditioning as it ignores this internal

organization"[82]. Esto es una razón más para aplicar métodos cuantitativos a la selección de objetos lexicográficos.

Este trabajo ha hecho evidente que a partir de una muestra, los índices de afijalidad y cliticidad propuestos en la introducción exhiben valores altos para afijos y palabras gramaticales y valores menores para todos los demás segmentos. Es decir que el carácter de ser elemento estructural o gramatical de un corpus es una propiedad que se puede estimar mediante los cálculos de entropía y economía inherentes a esos segmentos gramaticales. De este hecho se desprende que se está midiendo algo que se cuela entre los segmentos y que los caracteriza mucho más que sus significantes: sus relaciones con el resto de los segmentos del corpus. Cada objeto en un corpus tiene una cantidad de este algo a cada uno de sus lados. lo que sirve para medir algunas de sus características en el corpus (por ej., su cliticidad verdadera o su carácter gramatical total). A este algo lo podemos llamar glutinosidad.

Todo esto ha mostrado cuantitativamente varias cosas del español. Por ejemplo. que los prefijos difieren de los sufijos en que los segundos forman parte de un aparato morfosintáctico denso que los hace más aptos de ser descubiertos mediante los índices cuantitativos propuestos. De manera tal vez paralela, las formas proclíticas —si bien pocas— resultan más aptas de descubrirse mediante estos métodos que las enclíticas (sencillamente porque las formas enclíticas prototípicas del español se sufijan a la palabra gráfica. por lo que no queda mucho qué contemplar —tal vez algunos adjetivos muy gramaticales— como enclíticos gráficamente separados). El caso es que se abren muchas interrogantes en cuanto a los posibles resultados de estos procedimientos en otras lenguas, por ejemplo. si los prefijos y

[82]De Kock. *op. cit.* [82] 1978, p. 58.

los enclíticos se descubren más fácilmente que los sufijos y los proclíticos. Lo importante es que podemos inferir que la neutralidad de las técnicas aplicadas (al no ser éstas especiales a los fenómenos del español y a pesar de que las diferencias tipológicas —que otras lenguas puedan tener con respecto al español— resulten en patrones muy diferentes) las hace en cierta medida universales: si bien puede ser que no todas las lenguas utilicen los mismo recursos de las mismas maneras (por ej., los sufijos en la morfosintaxis), cabe esperar que todas sean sistemas en alguna medida entrópicos que hagan uso de diversas estrategias de economía. es decir, que entre sus segmentos más gramaticales haya algo de esa fuerza glutinosa que los caracteriza con respecto al corpus donde ocurran.

Si todo esto es cierto, se vuelve indispensable investigar todos los métodos que puedan medir estas propiedades. No se puede privilegiar uno o algunos métodos cuando sabemos que no son el fenómeno mismo que nos incumbe. No los podemos confundir. Los formalismos son nada más el andamiaje que nos acercan al edificio del lenguaje y nos ayudan a medirlo. Y creo que, si lo que queremos es simularlo, tampoco dejarán de ser más que un andamiaje.

Lo cierto es que, a pesar de que las computadoras se empezaron a aplicar al estudio y procesamiento del lenguaje y la literatura desde el final de la segunda guerra mundial, todavía no se han explorado todas las maneras en que se pueden aplicar en este campo. Como hemos visto, la reflexión lingüística tiene muchos caminos; aquí apenas se exploró uno, en el que sobre todo la curiosidad ha servido de guía principal. De esta curiosidad por identificar objetos lingüísticos y explorar los métodos para hacerlo, surgió la necesidad de medir las fuerzas imaginadas de afijalidad, cliticidad y —su generalización— glutinosidad. Es cierto que, como opinaría Javier Marías, los sacrosantos métodos cuantitativos involucrados

en una medición de estas propiedades lingüísticas implican la reducción y simplificación de la realidad al máximo. Sin embargo, no parece haber otra forma de conocerlas verdaderamente. especialmente si para esto optamos por seguir una de las reglas básicas de Galileo: medir todo lo medible e intentar hacer medible todo lo que aún no lo es[83].

[83] Citado por Bunge en *op. cit.* [25] 1967, p. 203.

Apéndice A

El *Corpus del Español Mexicano Contemporáneo*

En este apéndice se describe brevemente la versión del *CEMC* utilizada en esta investigación. Además de sus características generales, se incluyen apartados sobre su preprocesamiento, sus grafías o vocablos más frecuentes y las abreviaturas encontradas en él automáticamente.

A.1 Descripción

El *CEMC* es una colección de casi 1,000 textos[1] —de alrededor de 2,000 palabras cada uno (agrupadas en párrafos escogidos al azar)— que se originaron en toda la República Mexicana entre 1921 y 1974. Se trata tanto de obras escritas como de transcripciones de entrevistas grabadas. Están agrupados en 14 géneros (véase la tabla A.1) clasificados en lengua culta (literatura, periodismo, ciencias, técnicas, discursos políticos, religión y habla culta), subculta (literatura popular, habla media, lírica popular) y no-estándar (textos dialectales,

[1]En estudios anteriores, por presión de tiempo que obligó a cerrar la compilación del corpus antes de completarlo, se tuvo que utilizar una versión de 996 textos; Lara, art. cit. [?] lara1990bn *op. cit.* [85] 1990, p. 55.; Ham Chande, art. cit. [63] 1979 en Lara, Ham y García, *op. cit.* [89] 1979, p. 46. En esta investigación se trabajó con la versión de mil textos.

Tabla A.1: Géneros en el *CEMC*

nivel[a]	género	número de textos	número de palabras
lengua culta	Literatura	160	269 788
	Periodismo	176	299 775
	Ciencias	180	246 313
	Técnicas	102	202 716
	Discursos políticos	18	31 971
	Religión	12	21 277
	Habla culta	30	69 473
lengua sub-culta	Literatura popular	63	127 459
	Habla media	30	59 567
	Lírica popular	24	45 149
lengua no estándar	Textos dialectales	130	259 881
	Documentos antropológicos	33	68 376
	Jergas	20	34 839
	Habla popular	28	54 461
	totales	996	1 891 045

[a]Tabla basada en las tablas en Lara y Ham, art. cit. [88] 1974 y en Ham, "Del 1 al 100 en lexicografía" [63] de Lara, Ham y García, *op. cit.* [89] 1979, p. 47.

documentos antropológicos, jergas y habla popular). Esta subdivisión en géneros es una de las estrategias para garantizar la diversidad de textos y por lo tanto la representatividad estadística de este corpus de lengua española hablada en México (el proceso aleatorio de selección de párrafos es otra de las estrategias).

Cada línea del corpus tiene una clave de 9 dígitos entre arrobas ('@') que se refieren al documento original de donde se extrajo el texto. Los tres primeros se refieren al número del texto (entre 000 y 999), los tres segundos a la página y los tres últimos a la línea de la publicación original (@tttppplll@)[2]. Además, al final de cada línea se encuentran las anotaciones gramaticales aplicadas a las palabras del corpus. Los primeros dos dígitos (entre arrobas) representan el número de palabras de esa línea. los caracteres alfanuméricos que siguen son las marcas gramaticales (@dd@0123456789ABC). Más adelante, en la tabla A.8 se

[2]Véase García Hidalgo, "La formalización del analizador gramatical del DEM" art. cit. [50] 1979. en Lara. Ham y García, *op. cit.* [89] 1979, p. 107.

explica el significado de estas marcas.

A.2 Preprocesamiento

Algunas características del corpus, debidas sobre todo a las limitaciones del equipo computacional disponible cuando fue compilado, presentan ciertas dificultades en su procesamiento. Por ejemplo, las computadoras en que se almacenaron no contaban con algunos símbolos de gran importancia para la lengua española, tales como la 'ñ', las vocales acentuadas, la 'ü' y los signos de apertura de interrogación y exclamación ('¿', '¡').

En la tabla A.2 aparecen los caracteres que se utilizaron para representar esos símbolos. El problema es que, como podemos ver allí, esos símbolos cumplieron también con otras funciones. Así, los signos '+' y '/' (que, como se ve en la tabla A.2, se utilizaron para representar la 'ñ' y los acentos respectivamente) conservaron su función de operadores aritméticos (suma y división), incluso en textos no matemáticos (como '−' en @014039007@GRANDES CONQUISTADORES: JULIO@ CE/SAR@ + HERNA/N@ CORTE/S@ = PITO@ PE/REZ@): la 'ü' ni siquiera se tomó en cuenta[3]; los signos '?' y '!' sirvieron tanto para abrir como para cerrar interrogaciones y exclamaciones (esto es, no se distingue entre '¿' y '?' o entre '¡' y '!'); etc. Ni siquiera los caracteres menos ambiguos (véase la tabla A.3) están totalmente libres de estos problemas. Por ejemplo, el asterisco '*' también se usa algunas veces como separador de textos (@1800431180*) y, prácticamente todos los signos de puntuación se utilizan alguna

[3]Puede ser que a veces se haya utilizado el paréntesis derecho ')' después de la 'u' para representar 'ü'. En un sólo texto hay varias ocurrencias del segmento GU)ISA (para mi desconocido) que por el paréntesis podría tal vez ser 'güisa'; por ej. en @9660000220QUE/ HA SIDO DE MI BU+I DE MI GU)ISA, LA HAS VISTO O... O... TIENE ALGU/N MAJE.

Tabla A.2: Caracteres con varias funciones en el *CEMC*

caracter	funciones	lugar
'@'	marca de control	separa texto de clave de la línea y de las marcas gramaticales
	marca nombres propios	al final del nombre
'^'	marca palabra en mayúscula	al final de la palabra
'/'	acentúa vocal	después de la vocal
	fin de línea (versos)	entre palabras
	aritmética (división)	en fórmulas
'+'	'ñ'	en lugar de 'ñ'
	aritmética (suma)	en fórmulas
'-'	parentética	entre frases
	marca cambio de voz	diálogos
	guión	une palabras gráficas
	aritmética (resta)	en fórmulas
'.'	punto y seguido	entre oraciones
	punto final	entre párrafos
	punto de abreviatura	al final de abreviatura
	puntos suspensivos	entre puntos
'!'	'¡'. abre exclamación	al inicio. en lugar de '¡'
	'!'. cierra exclamación	al final
'?'	'¿'. abre pregunta	al inicio. en lugar de '¿'
	'?'. cierra pregunta	al final
'"'	abre cita	al inicio de cita
	cierra cita	al final de cita
	tipo de operador	en ciertas fórmulas
'''	enfatiza	alrededor de palabras
	abre cita	al inicio de cita
	cierra cita	al final de cita
	marca palabra incompleta	inicio o final de palabra

vez en alguna fórmula de algún tipo (por ej. ':', '"', ';' y '.' en `@388001189-90@POSTULADOS:`

`VPI) +: X X X " X ; + (X,Y) = X + Y`).

Por otra parte. varias líneas están repetidas, es decir. hay líneas consecutivas idénticas en todo menos en las marcas gramaticales. Algunas forman parte de letras de canciones, pero la mayoría son renglones de prosa repetidos con diferentes análisis gramaticales.

Para eliminar algunas de las ambigüedades de caracteres y las líneas duplicadas. se creó automáticamente. a partir del archivo que contiene el corpus original, el archivo `CEMC2.TXT`

Tabla A.3: Caracteres menos ambiguos en el *CEMC*

caracter	función	lugar
' ' ':', ':', ':'	puntuación (espacio en blanco) puntuación	marca de límites entre palabras entre palabras
'='	aritmética (igualdad)	en fórmulas
'*'	aritmética (multiplicación)	en fórmulas

que sirvió de base para sacar las listas de vocablos y construir las cadenas de Markov. Así, el archivo original tiene 219,315 líneas, mientras que el nuevo `CEMC2.TXT` conserva 219,141.

También se eliminaron las arrobas '@' que sirven para separar el texto de los datos asociados a cada renglón del corpus y las marcas gramaticales agregadas después de su recolección. Sólo se conservaron las arrobas que son marcas de nombres propios al interior del texto. Por otra parte, se dejaron las claves de referencia a los textos originales, pero sólo al interrumpirse la secuencia de las líneas originales. Las tablas A.4 (texto original) y A.5 (`CEMC2.TXT`) son fragmentos que muestran estos cambios. La segunda muestra al texto segmentado en los párrafos originales mediante las claves en un renglón separado y entre diagonales. De esta manera, ni las claves ni las marcas gramaticales (como se ve también eliminadas) interrumpen al texto. El texto ahora aparece en minúsculas, excepto las primeras letras de los nombres propios, que siguen marcados con una arroba al final. Las vocales con diagonal se convirtieron a vocales acentuadas, los signos de más '+' se cambiaron por eñes y los de exclamación (e interrogación) al principio de frase u oración se sustituyeron por sus versiones invertidas.

Debido a las ambigüedades ilustradas en la tabla A.2, aunque la mayor parte de los cambios se hizo automáticamente, muchos de éstos sólo se pudieron hacer manualmente. Así, por

Tabla A.4: Aspecto de un fragmento del *CEMC*

clave	texto	n[a] m[b]
@0000007003@DE PRONTO SE OYO/ UN DISPARO, EL PERRO LANZO/ UN GEMIDO SORDO Y NO	@14@41596
@0000007004@LADRO/ MA/S.	@2 @91
@0000013014@DEMETRIO@ DESPERTO/ SOBRESALTADO, VADEO/ EL RI/O Y TOMO/ LA VERTIENTE	@10@B9196
@0000013015@OPUESTA DEL CA+O/N. COMO HORMIGA ARRIERA ASCENDIO/ LA CRESTERI/A,	@9 @87818
@0000013016@CRISPADAS LAS MANOS EN LAS PE+AS Y RAMAZONES, CRISPADAS LAS PLANTAS	@11@86846
@0000013017@SOBRE LAS GUIJAS DE LA VEREDA.	@6 @46846
@0000020007@LOS FEDERALES GRITABAN A LOS ENEMIGOS, QUE, OCULTOS, QUIETOS Y	@10@68946
@0000020008@CALLADOS, SE CONTENTABAN CON SEGUIR HACIENDO GALA DE UNA PUNTERI/A QUE	@11@85949
@0000020009@YA LOS HABI/A HECHO FAMOSOS.	@5 @15998
@0000023012@DEMETRIO@ SIGUIO/ TIRANDO Y ADVIRTIENDO DEL GRAVE PELIGRO A LOS OTROS;	@11@B9939
@0000023013@PERO E/STOS NO REPARARON EN SU VOZ DESESPERADA SINO HASTA QUE SINTIERON	@12@35194
@0000023014@EL CHICOTEO DE LAS BALAS POR UNO DE LOS FLANCOS.	@10@68468
@0000027003@Y LOS SERRANOS, DESPUE/S DE ESTRECHARLES FUERTEMENTE LAS MANOS	@9 @36814
@0000027004@ENCALLECIDAS, EXCLAMABAN:	@2 @89
@0000027005@- !DIOS@ LOS BENDIGA! !DIOS@ LOS AYUDE Y LOS LLEVE POR BUEN CAMINO!...	@12@B59B5
@0000027006@AHORA VAN USTEDES; MA+ANA CORREREMOS TAMBIE/N NOSOTROS, HUYENDO DE LA	@10@19519
@0000027007@LEVA, PERSEGUIDOS POR ESTOS CONDENADOS DEL GOBIERNO, QUE NOS HAN	@10@89428
@0000027008@DECLARADO GUERRA A MUERTE A TODOS LOS POBRES; QUE NOS ROBAN NUESTROS	@12@98484
@0000027010@QUE QUEMAN NUESTRAS CASAS Y SE LLEVAN NUESTRAS MUJERES, Y QUE, POR FIN,	@13@09283
@0000027011@DONDE DAN CON UNO, ALLI/ LO ACABAN COMO SI FUERA PERRO DEL MAL.	@13@19451

[a]Número de palabras por renglón (entre arrobas).

[b]Marcas gramaticales. Hay tantas como palabras por renglón, pero por falta de espacio, sólo aparecen las primeras cinco. Estas marcas están explicadas en la tabla A.8.

ejemplo, todo signo de más '+' que no ocurrió entre caracteres alfabéticos e inmediatamente antes de una vocal no fue convertido a 'ñ', toda diagonal que no ocurrió junto a una vocal permaneció como diagonal. Pero, por otra parte, todas las 'u' con diéresis 'ü' que existen en el archivo **CEMC2.TXT** se agregaron manualmente (es decir, se examinaron todas las secuencias GUE y GUI para determinar por el contexto si pertenecían o no a palabras con diéresis[4]).

Por último, había en el corpus unos pocos caracteres del conjunto extendido ASCII (como se sabe, hay varios conjuntos extendidos) que, al no tener indicios de a cuál conjunto pertenecían, resultaban símbolos difíciles de interpretar (cuando posible). Estos símbolos fueron sustitui-

[4]Esto es, sin duda, un atrevimiento de mi parte, pero dada la naturaleza de esta investigación (sobre todo del primer capítulo), era necesario distinguir entre las formas con diéresis ('vergüenza', 'argüir', 'güera', etc.) y aquellas sin diéresis ('guerra', 'distingue', 'guitarra', etc.).

Tabla A.5: Aspecto de un fragmento del archivo **CEMC2.TXT**

texto
\000007003\
de pronto se oyó un disparo, el perro lanzó un gemido sordo y no ladró más.
\000013014\
Demetrio@ despertó sobresaltado, vadeó el río y tomó la vertiente opuesta del cañón. como hormiga arriera ascendió la crestería, crispadas las manos en las peñas y ramazones, crispadas las plantas sobre las guijas de la vereda.
\000020007\
los federales gritaban a los enemigos, que, ocultos, quietos y callados, se contentaban con seguir haciendo gala de una puntería que ya los había hecho famosos.
\000023012\
Demetrio@ siguió tirando y advirtiendo del grave peligro a los otros; pero éstos no repararon en su voz desesperada sino hasta que sintieron el chicoteo de las balas por uno de los flancos.
\000027003\
y los serranos, después de estrecharles fuertemente las manos encallecidas, exclamaban:
- ¡Dios@ los bendiga! ¡Dios@ los ayude y los lleve por buen camino!... ahora van ustedes; mañana correremos también nosotros, huyendo de la leva, perseguidos por estos condenados del gobierno, que nos han declarado guerra a muerte a todos los pobres; que nos roban nuestros puercos, nuestras gallinitas y hasta el maicito que tenemos para comer; que queman nuestras casas y se llevan nuestras mujeres, y que, por fin, donde dan con uno, allí lo acaban como si fuera perro del mal.

dos por uno sólo '□' (véase el núm. 42 de la tabla A.6) para disminuir el número de caracteres del corpus.

Los caracteres del corpus

Una vez creado el archivo **CEMC2.TXT** (el corpus con modificaciones), se procedió a hacer un inventario de los caracteres y sus frecuencias en el corpus. En la tabla A.6 se listan los caracteres encontrados por orden de aparición en el corpus. Luego, se hicieron varias listas de vocablos. Una sin hacer ninguna modificación a los caracteres, otra eliminando los acentos, una más sin omitir los acentos en la última vocal y haciendo cambios para reflejar

Tabla A.6: Frecuencias y porcentajes de caracteres en CEMC2.TXT

	caracter[a]	frecuencia	porcentaje		caracter	frecuencia	porcentaje
1.	'd'	441669	3.7%	28.	':'	9511	0.08%
2.	'e'	1263477	11%	29.	'x'	18015	0.15%
3.	' '[b]	2052833	17%	30.	';'	9010	0.076%
4.	'p'	247362	2.1%	31.	'-'	26846	0.23%
5.	'r'	589488	4.9%	32.	'¡'	4807	0.04%
6.	'o'	872198	7.3%	33.	'!'	5180	0.043%
7.	'n'	646607	5.4%	34.	'¿'	10894	0.091%
8.	't'	416712	3.5%	35.	'?'	11696	0.098%
9.	's'	713549	6%	36.	'"'	19224	0.16%
10.	'y'	102277	0.86%	37.	'ü'	322	0.0027%
11.	'u'	391955	3.3%	38.	'9'[c]	40795	0.34%
12.	'i'	613081	5.1%	39.	'('	7073	0.059%
13.	'a'	1156514	9.7%	40.	')'	8135	0.068%
14.	','	155623	1.3%	41.	'k'	2561	0.021%
15.	'l'	507592	4.3%	42.	'□'[d]	53	0.00044%
16.	'z'	34190	0.29%	43.	'w'	1052	0.0088%
17.	'g'	101609	0.85%	44.	'+'	383	0.0032%
18.	'm'	276155	2.3%	45.	'='	346	0.0029%
19.	'.'	217931	1.8%	46.	'/'	633	0.0053%
20.	'b'	117713	0.99%	47.	'&'	8	6.7e-05%
21.	'v'	94060	0.79%	48.	'¿'	1184	0.0099%
22.	'c'	414090	3.5%	49.	'"'	1349	0.011%
23.	'ñ'	15972	0.13%	50.	'%'	610	0.0051%
24.	'h'	86126	0.72%	51.	'$'	215	0.0018%
25.	'j'	43002	0.36%	52.	'*'	235	0.002%
26.	'f'	67080	0.56%	53.	'¡'	1126	0.0094%
27.	'q'	103275	0.87%	54.	'#'	38	0.00032%
					Total	11923441	100%

[a]Las vocales acentuadas se contaron como vocales no acentuadas.
[b]Espacio en blanco.
[c]Cualquier dígito (entre 0 y 9).
[d]Símbolos irreconocibles.

la correspondencia entre letras y fonemas del español (véase la tabla A.7 que lista estas correspondencias) y otra última con las modificaciones, pero omitiendo todos los acentos.

Los procedimientos descritos en el primer capítulo se aplicaron a las cuatro listas de vocablos. Aunque todavía merecen analizarse, las diferencias entre los resultados de las cuatro listas no fueron muy grandes. Los datos presentados en este trabajo, como se explicó en el primer capítulo, son resultado del procesamiento de la lista de vocablos con acentos en

Tabla A.7: Modificaciones a caracteres para reflejar correspondencia entre grafemas y fonemas

modificaciones	fonema	contextos
'v' → 'b'	[b]	todos
'z', 'c' → 's	[s]	toda 'z'; 'ce', 'ci'
'c', 'qu' → 'k'	[k]	'ca', 'que', 'qui', 'co', 'cu'
'ch' → 'č'	[č]	todos
'g' → 'g'	[γ]	'ga', 'go', 'gu'
'gu' → 'g'	[γ]	'gue', 'gui'
'g' → 'j'	[h]	'ge', 'gi'
'h' → ξ	-	todos
'y' → 'i'	[i]	fin de sílaba, después de vocal ('ay', 'ey', ...).
'y', 'll' → 'y'	[y]	principio de sílaba, antes de vocal
'rr' → 'r̄'	[r̄]	todos
'r' → 'r̄'	[r̄]	principio de palabra; o después de sílaba que termina en 'n', 'l', 's' or 'b'.
'r' → 'r'	[r]	entre vocales.

la última vocal y con las modificaciones de la tabla A.7.

A.3 Las marcas gramaticales del *CEMC*

En la tabla A.8 se listan las marcas gramaticales aplicadas al *CEMC*. Las marcas del 0 al 9 se describen en García Hidalgo[5]. Las explicaciones de las marcas A, B y C son observaciones mías. Como lo explica Isabel García, las marcas se aplicaron al corpus mediante el analizador sintáctico construido por ella, procedimiento que mano calificada completó después por inspección.

A.4 Formas más frecuentes

Como se sabe, el objetivo de la construcción del *CEMC* fue la recolección de los voca-

[5]García, art. cit. [50] 1979, p. 91; corresponden a las categorías gramaticales del conjunto *G*.

Tabla A.8: Marcas gramaticales del *CEMC*

marca	tipo de palabra
0	ambigua
1	adverbio
2	adjetivo
3	conjunción
4	preposición
5	pronombre
6	artículo
7	contracción
8	nominal
9	verbo
A	apoyos conversacionales[a],
B	nombres propios[b]
C	otros[c]

[a]Por ejemplo, ¿VERDAD?, BUENO, ¿EH?, EH, AH, ÉJELE, UAAO, BAH, JIJI, JEJE, AY, HUM, etc., así como interjecciones (¡VAYA!, ¡FUCHI!, ¡LÁSTIMA!, etc.).

[b]Por ej., BENITO@ JUÁREZ@, MIXCOAC@.

[c]Números (por ej., 1918, 7:30, 17:45), palabras con mayúscula (por ej., GOBERNACIÓN^, SECRETARÍA^), errores (BASTIM#ENTO, QU#'L), etc.

blos más representativos del español mexicano que deben incluirse en la nomenclatura del Diccionario del Español de México. Uno de los criterios es, naturalmente, la frecuencia que, sin embargo, sufre de importantes fluctuaciones según, por ejemplo, el género de los textos. Para solucionar este problema, se aplicaron varias correcciones[6].

A continuación se presentan brevemente algunos resultados de aquella investigación (los cien lemas más comunes en el corpus) y las formas más frecuentes según este trabajo. Por último, se describe el procedimiento para determinar las abreviaturas más frecuentes del corpus que se aplicó en este trabajo tanto para desambiguar el uso del punto '.', como para evitar que estas formas fueran tomadas como vocablos plenos del corpus.

[6]Que se presentaron en el segundo capítulo de esta tesis (a partir de la página 38) y que, como se menciona allí, se describen en Lara y Ham Chande, art. cit. [88] 1974.

A.4.1 Vocablos

En este apartado se presentan los tipos de palabras más frecuentes del *CEMC*. En la tabla A.9 se reproduce la lista de formas lematizadas más frecuentes compilada por Carlos Villanueva[7]. Las formas aparecen en el orden de la frecuencia corregida que se calculó para tomar en cuenta la dispersión de los vocablos entre los diferentes géneros del corpus[8]. En contraste, en la tabla A.10 que aparece más adelante (a partir de la página 365) se listan las casi doscientas formas más frecuentes (sin lematizar), según las cuentas en esta investigación. Las diferencias son producto de, además de la lematización, las modificaciones efectuadas a las formas en este experimento (descritas arriba en la tabla A.7) que causan que varios vocablos sean homónimos (y aparezcan en una misma forma), mientras que en la tabla de los lemas del *CEMC* las formas homónimas aparecen separadas (art. 'la' es el núm. 1, mientras que el pron. 'la' el 46).

A.4.2 Abreviaturas

Esta sección se ocupa de las abreviaturas que se pueden identificar automáticamente en el *CEMC*. Como se vio en la sección sobre el fichador (*tokenizer*) del primer capítulo, decidir si un punto separa períodos o es marca de una abreviatura no es tarea fácil. Por esa razón, se construyó un programa especial para compilar una lista de segmentos que funcionen como abreviaturas. La mayoría de las abreviaturas que aparecen en la tabla A.11

[7]Publicada en Ham Chande. "Del 1 al 100 en lexicografía" art. cit. [63] 1979, pp. 54-55.

[8]Como ya se dijo, la descripción detallada de este índice está en Lara y Ham Chande, art. cit. [88] 1974, p. 37.

fueron seleccionadas automáticamente según los criterios descritos a continuación.

Primero, se contaron las veces que los segmentos ocurrieron como abreviaturas probables, es decir, cuando ocurrieron seguidas de un punto (especialmente si inmediatamente después de dicho punto ocurre otro signo de puntuación, ',', ':', ';', '?', etc.). Luego se eliminaron las formas más largas (las de más de cinco caracteres). Por último, se consideró la estructura de las formas al hacer un examen automático para determinar si terminaban en consonantes con las que suelen o no terminar las palabras del español, si contenían o no vocales y si estaban rodeadas de consonantes a la manera de los patrones silábicos conocidos del español.

De esta manera, al tomar el número de veces que un segmento (de menos de seis caracteres organizados de una manera peculiar con respecto al patrón silábico de la lengua examinada) ocurre como abreviatura (es decir, cuando va seguida de un punto) y dividirlo entre su frecuencia total en el corpus, se obtiene una estimación de la probabilidad que tiene dicho segmento de ser una abreviatura. En la tabla A.11 se listan todos los segmentos del *CEMC* con estas características. La probabilidad es el cociente de las columnas de frecuencias, esto es, de la frecuencia como abreviatura sobre la frecuencia total del segmento.

Las pocas abreviaturas que están marcadas con un asterisco al final del renglón no fueron resultado de este procedimiento, sino se agregaron después porque, aunque sí ocurrieron en el corpus y se trata de abreviaturas muy conocidas, lo hicieron muy pocas veces o su probabilidad fue muy baja; o porque, aunque no ocurrieron en el corpus, se trata de letras aisladas que bien podrían ocurrir en córpora futuros.

Tabla A.9: Los lemas más frecuentes del *CEMC* (frecuencias corregidas)

núm.	lema	categoría	fr. corregida[a]	dispersión norm.
001	la	art.	85919.13	0.9780
002	el	art.	78942.85	0.9806
003	de	prep.	63088.75	0.9941
004	y	conj.	59255.11	0.9962
005	que	pron.	58364.00	0.9960
006	en	prep.	48228.00	0.9780
007	a	prep.	44569.77	0.9946
008	se	pron.	33442.49	0.9922
009	no	adv.	28670.65	0.9101
010	ser	v.	24428.99	0.9802
011	un	art.	20561.20	0.9958
012	por	prep.	19694.88	0.9915
013	con	prep.	18660.50	0.9938
014	su	adj.	17233.52	0.9665
015	una	art.	15597.41	0.9939
016	haber	v.	14373.69	0.9830
017	para	prep.	13951.97	0.9849
018	al	contr.	11028.71	0.9839
019	estar	v.	10940.95	0.9366
020	como	adv.	10753.78	0.9940
021	tener	v.	9402.55	0.9572
022	le	pron.	9192.13	0.8707
023	hacer	v.	8607.80	0.9750
024	ya	adv.	8238.52	0.8376
025	o	conj.	7989.34	0.9542
026	pero	conj.	7706.24	0.9228
027	decir	v.	7612.18	0.8555
028	que	conj.	7565.24	0.9477
029	lo	art.	7335.91	0.9942
030	me	pron.	6928.90	0.6607
031	más	adv.	6877.20	0.9874
032	poder	v.	6200.40	0.9762
033	este	adj.	6180.97	0.8639
034	ir	v.	5621.48	0.7954
035	lo	pron.	5549.14	0.9060
036	sí	adv.	4976.58	0.5701
037	ver	v.	4891.83	0.8775
038	dar	v.	4876.03	0.9661
039	cuando	adv.	4751.19	0.9679
040	muy	adv.	4503.17	0.9156
041	yo	pron.	4390.14	0.6210
042	porque	conj.	4202.55	0.8255
043	él	pron.	4101.76	0.9085
044	mi	adj.	4046.84	0.7012
045	pues	conj.	4005.46	0.6731

[a] *Korregierte Frequenz*; véase Lara y Ham Chande, art. cit. [88] 1974. p. 37.

Tabla A.9 (continuación):
Los lemas más frecuentes del *CEMC* (frecuencias corregidas)

núm.	lema	categoría	fr. corregida	dispersión norm.
046	la	pron.	3927.33	0.9212
047	así	adv.	3739.74	0.8985
048	esta	adj.	3530.77	0.9447
049	todo	adj.	3428.97	0.9924
050	también	adv.	3388.18	0.9682
051	vez	s.	3152.41	0.9730
052	nos	pron.	3146.58	0.9277
053	año	s.	3102.49	0.8954
054	saber	v.	3085.00	0.8954
055	sin	prep.	2999.75	0.9401
056	hasta	prep.	2979.90	0.9820
057	querer	v.	2916.40	0.8204
058	deber	v.	2893.21	0.9033
059	todo	pron.	2892.46	0.9212
060	aquí	adv.	2818.17	0.6953
061	uno	pron.	2655.45	0.8558
062	día	s.	2623.75	0.9414
063	eso	pron.	2526.39	0.7509
064	qué	pron.	2522.64	0.7919
065	ella	pron.	2410.23	0.9212
066	sobre	prep.	2410.01	0.8941
067	bien	adv.	2361.08	0.9442
068	llegar	v.	2348.04	0.9611
069	más	adj.	2341.78	0.9494
070	donde	adv.	2274.15	0.9800
071	entre	prep.	2268.93	0.9226
072	ni	conj.	2249.99	0.9404
073	otra	adj.	2242.32	0.9899
074	entonces	adv.	2229.34	0.7503
075	esa	adj	2227.29	0.9542
076	llevar	v.	2185.06	0.9507
077	poner	v.	2129.97	0.9176
078	parte	s.	2116.71	0.9439
079	te	pron.	2048.26	0.5993
080	tiempo	s.	2047.34	0.9899
081	dos	s.	1995.51	0.9882
082	después	adv.	1988.68	0.9770
083	dejar	v.	1982.57	0.9444
084	desde	prep.	1886.32	0.9875
085	hombre	s.	1877.77	0.9434
086	ese	adj.	1869.75	0.9437
087	cada	adj.	1826.01	0.9689
088	venir	v.	1786.61	0.8515
089	quedar	v.	1786.50	0.9590
090	ahora	adv.	1785.13	0.9290

Tabla A.9 (continuación):
Los lemas más frecuentes del *CEMC* (frecuencias corregidas)

núm.	lema	categoría	fr. corregida	dispersión norm.
091	esto	pron.	1758.29	0.9706
092	pasar	v.	1756.64	0.9207
093	nada	pron.	1722.46	0.7933
094	siempre	adv.	1610.75	0.9659
095	vida	s.	1560.79	0.9188
096	casa	s.	1546.17	0.8512
097	sólo	adv.	1531.92	0.8931
098	tomar	v.	1505.79	0.9801
099	forma	s.	1501.64	0.8783
100	trabajo	s.	1495.21	0.9693

Tabla A.10: Las formas más frecuentes en CEMC2.TXT

núm.	forma[a]	fr. absoluta	grafías
001	[de]	118879	de, dé
002	[la]	75963	la (art., pron.)
003	[ke]	71801	que, qué
004	[i]	65356	y
005	[a]	58983	a, ha
006	[el]	56771	el, él
007	[en]	51951	en
008	[se]	35658	sé, se
009	[los]	31985	los (art., pron.)
010	[no]	31362	no
011	[las]	21364	las (art., pron.)
012	[un]	20277	un
013	[por]	20054	por
014	[es]	19681	es
015	[del]	19168	del
016	[kon]	19072	con
017	[una]	16669	una (art., pron.)
018	[o]	15422	o
019	[para]	14796	para
020	[si]	14716	sí, si
021	[lo]	14093	lo (art., pron.)
022	[su]	12954	su
023	[komo]	12697	cómo, como (adv., conj., v.)
024	[al]	11793	al
025	[e]	10829	he, e
026	[me]	10080	me
027	[mas]	9945	más, mas
028	[ya]	9547	ya
029	[le]	8470	le
030	[pero]	8180	pero
031	[ai]	7734	hay, ahí, ay
032	[este]	7683	éste, este
033	[esta]	7418	ésta, esta
034	[yo]	6669	yo
035	[mi]	5894	mí, mi (adj., sust.)
036	[pues]	5853	pues
037	[sus]	5386	sus
038	[mui]	4915	muy
039	[porke]	4886	porque
040	[kuando]	4810	cuándo, cuando
041	[son]	4494	son (v., sust.)
042	[todo]	4083	todo (adj., pron.)
043	[asi]	4041	así
044	[te]	4035	té, te
045	[aki]	3858	aquí

[a]No se tomaron en cuenta los acentos, se eliminaron las abreviaturas y se aplicaron las modificaciones descritas en la tabla A.7.

Tabla A.10 (continuación):
Las formas más frecuentes en CEMC2.TXT

núm.	forma	fr. absoluta	grafías
046	[tambien]	3474	también
047	[nos]	3422	nos
048	[dos]	3294	dos (sust., adj.)
049	[sin]	3279	sin
050	[eso]	3087	eso
051	[tiene]	3068	tiene
052	[fue]	3045	fue
053	[asta]	2992	hasta
054	[uno]	2963	uno (sust. (núm), pron.)
055	[ser]	2946	ser
056	[sobre]	2786	sobre
057	[entonses]	2768	entonces
058	[era]	2728	era
059	[nada]	2675	nada (sust., adv.)
060	[todos]	2632	todos (adj., pron.)
061	[entre]	2611	entre (prep., v.)
062	[bien]	2561	bien
063	[tu]	2558	tú. tu
064	[puede]	2462	puede
065	[ni]	2406	ni
066	[ese]	2361	ése. ese
067	[an]	2345	han
068	[donde]	2310	dónde. donde
069	[bes]	2306	vez. ves
070	[bueno]	2159	bueno
071	[solo]	2146	sólo. solo
072	[les]	2125	les
073	[años]	2115	años
074	[ba]	2076	va
075	[otra]	2060	otra (adj., pron.)
076	[ase]	2035	hace
077	[despues]	2021	después
078	[eya]	2020	ella
079	[abia]	1981	había
080	[otro]	1977	otro (adj., pron.)
081	[parte]	1899	parte
082	[desde]	1897	desde
083	[kada]	1897	cada
084	[tiempo]	1896	tiempo
085	[esa]	1889	esa (adj., pron.)
086	[aora]	1884	ahora
087	[luego]	1828	luego
088	[mismo]	1797	mismo
089	[dia]	1777	día
090	[dise]	1759	dice

Tabla A.10 (continuación):
Las formas más frecuentes en CEMC2.TXT

núm.	forma	fr. absoluta	grafías
091	[mučo]	1734	mucho
092	[bida]	1700	vida
093	[aser]	1688	hacer
094	[kien]	1664	quién, quien
095	[tan]	1660	tan
096	[as]	1648	has, haz, as
097	[aya]	1647	halla, haya (v., sust.)
098	[siempre]	1635	siempre
099	[tres]	1634	tres (sust., adj.)
100	[sea]	1623	sea
101	[esto]	1593	esto
102	[kasa]	1557	casa (sust., v.), caza (sust. v.)
103	[estan]	1535	están
104	[forma]	1530	forma
105	[otros]	1500	otros
106	[digo]	1488	digo
107	[usted]	1465	usted
108	[ber]	1455	ver
109	[eyos]	1440	ellos
110	[dijo]	1421	dijo
111	[agua]	1415	agua
112	[ombre]	1415	hombre
113	[poko]	1415	poco
114	[tanto]	1415	tanto (adj., pron., adv.)
115	[menos]	1395	menos
116	[trabajo]	1377	trabajo
117	[desir]	1353	decir
118	[todas]	1347	todas (adj., pron.)
119	[sino]	1332	sino
120	[debe]	1329	debe
121	[estaba]	1315	estaba
122	[beses]	1309	veces
123	[antes]	1300	antes
124	[asia]	1281	hacia, hacía
125	[gran]	1278	gran
126	[tienen]	1274	tienen
127	[mayor]	1254	mayor
128	[estos]	1243	estos (adj., pron.)
129	[pos]	1241	pos (pues)
130	[dias]	1236	días
131	[año]	1227	año
132	[tal]	1208	tal
133	[boi]	1193	voy
134	[toda]	1190	toda (adj., pron.)
135	[estado]	1186	estado (sust., v.)

Tabla A.10 (continuación):
Las formas más frecuentes en CEMC2.TXT

núm.	forma	fr. absoluta	grafías
136	[da]	1152	da
137	[unos]	1150	unos (art., pron.)
138	[kaso]	1145	caso (sust., v.), cazo (sust., v.)
139	[ps]	1132	ps (pues)
140	[kasi]	1128	casi
141	[estas]	1111	éstas, estas
142	[mejor]	1105	mejor (adj. y v.)
143	[algo]	1099	algo (adv., pron.)
144	[kosas]	1085	cosas
145	[otras]	1083	otras (adj., pron.)
146	[kosa]	1082	cosa
147	[kual]	1078	cuál, cual
148	[ejemplo]	1072	ejemplo
149	[tengo]	1072	tengo
150	[medio]	1071	medio
151	[sido]	1069	sido
152	[nosotros]	1051	nosotros
153	[tenia]	1051	tenía, tenia
154	[lugar]	1050	lugar
155	[nunka]	1045	nunca
156	[jeneral]	1038	general (adj., sust.)
157	[mujer]	1036	mujer
158	[durante]	1030	durante
159	[ečo]	1022	hecho (v., sust.), echo, echó
160	[mundo]	999	mundo
161	[ora]	998	hora, ora
162	[señor]	998	señor
163	[pesos]	965	pesos
164	[nasional]	954	nacional
165	[pueden]	954	pueden
166	[mučos]	942	muchos (adj., pron.)
167	[jente]	937	gente
168	[mis]	933	mis
169	[aunke]	927	aunque
170	[pais]	910	país
171	[ir]	904	ir
172	[ban]	903	van
173	[berdad]	902	verdad
173	[dentro]	890	dentro
174	[misma]	887	misma
175	[esos]	864	ésos, esos
176	[kreo]	861	creo, creó
177	[pa]	860	pa' (papá, para)
178	[mil]	856	mil
179	[fueron]	852	fueron
180	[nuestro]	852	nuestro

Tabla A.10 (continuación):
Las formas más frecuentes en CEMC2.TXT

núm.	forma	fr. absoluta	grafías
181	[oi]	852	hoy, oí
182	[primera]	851	primera
183	[ayi]	848	allí
184	[grandes]	849	grandes
185	[primer]	838	primer
186	[manera]	832	manera
187	[kuatro]	828	cuatro
188	[algunos]	826	algunos

Tabla A.11: Abreviaturas en el *CEMC*

núm.	abrev.	fr. abrev.	fr. total	prob.	
001	A	805	46185	0.0174299	
002	Á	1	12	0.0833333	*
003	ÁC	57	57	1	
004	ÁCS	9	9	1	
005	ACT	4	5	0.8	
006	AGS	2	2	1	
007	ALT	2	2	1	
008	AMP	4	5	0.8	
009	AMPS	2	2	1	
010	ANH	3	3	1	
011	APROX	1	1	1	*
012	ART	27	38	0.710526	
013	ARTS	3	3	1	
014	B	657	1079	0.608897	
015	BDH	2	3	0.666667	
016	C	178	664	0.268072	
017	CAP	4	5	0.8	
018	CARB	2	2	1	
019	CC	19	34	0.558824	
020	CF	10	14	0.714286	
021	CFA	3	3	1	
022	CFR	3	3	1	
023	CIT	4	4	1	
024	CMS	20	21	0.952381	
025	COAH	9	9	1	
026	COEFS	1	1	1	*
027	CONC	9	9	1	
028	COR	7	11	0.636364	
029	CTMS	2	2	1	
030	CUC	3	5	0.6	
031	CÍA	5	6	0.833333	
032	CH	1	15	0.0666667	*
033	CHIH	7	7	1	
034	CHIS	4	5	0.8	
035	D	166	367	0.452316	
036	DGO	6	6	1	
037	DIL	2	3	0.666667	
038	DR	44	52	0.846154	
039	DÉC	2	2	1	
040	E	112	1456	0.0769231	
041	EB	9	9	1	
042	EC	5	7	0.714286	
043	EF	4	5	0.8	
044	EJ	6	8	0.75	
045	EJEM	3	4	0.75	

Tabla A.11 (continuación):
Abreviaturas en el *CEMC*

núm.	abrev.	fr. abrev.	fr. total	prob.	
046	ENC	937	938	0.998934	
047	ENF	7	7	1	
048	ENT	2	2	1	
049	ESP	11	12	0.916667	
050	ETC	548	603	0.908789	
051	F	160	352	0.454545	
052	FCO	2	2	1	
053	FF	3	3	1	
054	FIG	310	319	0.971787	
055	FIGS	7	7	1	
056	G	147	400	0.3675	
057	GAL	2	2	1	
058	GEN	2	3	0.666667	
059	GR	19	21	0.904762	
060	GRAL	3	5	0.6	
061	GRMS	3	3	1	
062	GRO	4	4	1	
063	GRS	2	2	1	
064	GTO	6	6	1	
065	H	70	274	0.255474	
066	HA	27	4134	0.0065312	*
067	HEMSL	2	2	1	
068	HGO	4	5	0.8	
069	HS	4	6	0.666667	
070	HV	5	8	0.625	
071	I	65	249	0.261044	
072	IBID	2	3	0.666667	
073	INC	5	6	0.833333	
074	INF	1779	1787	0.995523	
075	ING	10	11	0.909091	
076	INST	3	3	1	
077	J	113	260	0.434615	
078	JAL	17	19	0.894737	
079	JR	17	28	0.607143	
080	K	20	64	0.3125	
081	KG	25	114	0.219298	*
082	KGS	9	9	1	
083	KPH	2	3	0.666667	
084	L	87	475	0.183158	
085	LC	3	3	1	
086	LIB	6	6	1	
087	LIC	31	34	0.911765	
088	LTS	2	2	1	
089	LÁM	9	9	1	
090	M	248	658	0.3769	

Tabla A.11 (continuación):
Abreviaturas en el *CEMC*

núm.	abrev.	fr. abrev.	fr. total	prob.	
091	MED	5	6	0.833333	
092	MET	3	4	0.75	
093	MGR	2	3	0.666667	
094	MICH	8	8	1	
095	ML	60	147	0.408163	*
096	MM	58	264	0.219697	*
097	MME	6	6	1	
098	MOR	6	7	0.857143	
099	MR	12	14	0.857143	
100	MTS	21	24	0.875	
101	MÁX	4	4	1	
102	MÉX	8	8	1	
103	N	55	235	0.234043	
104	NOV	8	9	0.888889	
105	NÚM	9	9	1	
106	NÚMS	2	2	1	
107	O	44	8330	0.00528211	
108	OAX	8	8	1	
109	P	156	320	0.4875	
110	PAG	3	3	1	
111	PREP	41	41	1	
112	PROF	9	10	0.9	
113	PÁG	36	38	0.947368	
114	PÁGS	3	3	1	
115	Q	6	26	0.230769	
116	QRO	2	2	1	
117	R	164	347	0.472622	
118	ROM	7	7	1	
119	S	128	257	0.498054	
120	SEC	3	5	0.6	
121	SEG	8	14	0.571429	
122	SPP	3	4	0.75	
123	T	34	151	0.225166	
124	TAMPS	2	2	1	
125	TAMS	2	2	1	
126	TLAX	2	2	1	
127	U	15	166	0.0903614	
128	UD	8	8	1	
129	UDS	2	3	0.666667	
130	UHM	15	16	0.9375	
131	UNF	2	2	1	
132	US	2	3	0.666667	
133	V	94	208	0.451923	
134	VD	4	5	0.8	
135	VEL	4	6	0.666667	

Tabla A.11 (continuación):
Abreviaturas en el *CEMC*

núm.	abrev.	fr. abrev.	fr. total	prob.
136	VOL	2	3	0.666667
137	VS	9	10	0.9
138	W	24	97	0.247423
139	X	37	439	0.0842825
140	Y	21	60421	0.000347561
141	YUC	3	3	1
142	Z	17	29	0.586207
143	ZAC	2	2	1
144	ZL	2	2	1

Apéndice B

Muestra aletoria de vocablos analizados

En este apéndice se consignan los vocablos del *CEMC* escogidos al azar para comparar los índices de segmentación automática examinados en el primer capítulo. La información principal aparece esencialmente en cuatro tablas. La tabla B.1 contiene los vocablos que formaban parte de la muestra, pero que fueron omitidos por inanalizables o de análisis dudoso, al tratarse de términos extranjeros o de morfología antigua. La tabla B.3 sirve para identificar los índices en cuestión con los signos que encabezan las columnas de la tabla B.4, la cual contiene los 845 vocablos de la muestra y sus análisis. Los resultados de los conteos de aciertos por índice aparecen en la tabla B.2.

La muestra original contenía 851 vocablos de cinco o más caracteres (representativos de fonemas, según la tabla A.7 del primer apéndice) cada uno. Pero se eliminaron aquellos que aparecen en la tabla B.1 debido a lo cuestionable de sus posibles segmentaciones. Allí hay préstamos ('bistec', 'konboi' y 'nokaut') cuya morfología es ajena a la española. Los restantes son palabras de origen grecolatino cuya morfología resulta un tanto distante de la española:

Tabla B.1: Vocablos de la muestra omitidos

	vocablo	sufs.	segmentaciones
846	bistec	?	bi+ste)]}',;.:c
847	impetu	?	impe].:.:t)+}u
848	juebes	?	ju)}e]'.b+e,::s
849	konboi	?	kon.:b.:o)+]}'i
850	metropoli	?	metropol)+}.;.:i
851	nokaut	?	no,;.:kau)+}t

'im-pet-u' (<'pedir'), 'juebes' (<*dies Jovis*) y 'metropoli', de los que tal vez cabría argüir que *~u*, *~es* e *~i* podrían considerarse afijos, pero no exactamente de la lengua española.

Los resultados del análisis de la muestra se presentaron en la tabla 2.7 del primer capítulo. En la tabla B.2, éstos se repiten junto con las cuentas de otros índices que no fueron tan pertinentes en ese capítulo. La mayoría se examinaron allí con detalle. El menos explicado fue el de substracción de entropías, el cual sencillamente se trata de la resta de los valores de entropía en una segmentación dada (la calculada de derecha a izquierda menos la de izquierda a derecha)[1]. El número de cuadros por cada sufijo alternante no se mencionó en

Tabla B.2: Resultados del análisis de la muestra

índice	aciertos	fracción	porcentaje
economía o entropía	807	0.955029586	95.50%
afijalidad	764	0.904142012	90.41%
cuadros	737	0.872189349	87.22%
entropía	730	0.863905325	86.39%
cuadros por cada sufijo alternante	705	0.834319527	83.43%
economía	669	0.791715976	79.17%
coeficiente de Yule	609	0.720710059	72.07%
información mutua	604	0.714792899	71.48%
prueba de χ^2	583	0.689940828	68.99%
razón de semejanza	582	0.688757396	68.88%
substracción de entropías	272	0.321893491	32.19%

[1]Dado el bajísimo porcentaje de aciertos de este índice (pensado tempranamente en este trabajo), es pertinente notar que se trata de las entropías de dos segmentos yuxtapuestos. De volverse a aplicar la misma idea, convendría restar las entropías de las fronteras de un sólo segmento a la vez (de la izquierda o derecha de cada segmento) y no las entropías de dos segmentos colindantes (hacia la derecha del uno y hacia la izquierda del otro) como se hizo al inicio de esta investigación.

el primer capítulo. Se trata simplemente del cociente de la cantidad de cuadros atestiguada en una segmentación entre el número de segmentos que se detectaron allí como alternantes del supuesto afijo. El último índice de esta tabla (véase el último renglón) es la cuenta de las veces en que cualquiera de los índices de economía o entropía (uno o ambos) acertó en la predicción de una segmentación morfológica (sin contar el índice de cuadros).

En cuanto a la tabla B.4, la principal de este apéndice, la primera columna contiene los vocablos. La segunda el sufijo o la cadena de sufijos observada (como se apuntó en el primer capítulo, los prefijos no se tomaron en cuenta). Los vocablos están agrupados según estos sufijos o cadenas de sufijos (por ej. todos los que terminan con el sufijo *~a* aparecen al principio). Nótese que algunos contienen más sufijos que los consignados en esta columna, pero no se tomaron en cuenta porque ninguno de los valores atestiguó un valor máximo entre esos sufijos y sus bases (recuérdese que sólo se tomó el más alto para cada vocablo). Por ejemplo, el vocablo *arkeolojika* (núm. 13), para el que se puede proponer la cadena de afijos *~ik.a*, aparece en el grupo de los vocablos con el sufijo *~a*, porque ningún índice propuso una segmentación entre *arkeoloj~* e *~ika*. Por otra parte, hay algunos que, para ser agrupados, se les especifican más sufijos que los que los índices propusieron, por ej. *bibiendo* (núm. 193).

La tercera columna contiene los vocablos marcados por los índices de segmentación examinados en el primer capítulo (en la tabla B.3 aparecen las correspondencias de índices con los signos usados para marcar las segmentaciones). Cada signo o marca corresponde a los valores más altos de cada índice de segmentación para cada vocablo. Las siguientes columnas representan a cada uno de los índices. Los aciertos aparecen señalados con un '1'. Cuando el

Tabla B.3: Correspondencias de signos para la tabla B.4

signo	índice
]	cuadros
)	economía
⟩	entropía
'	cuadros por cada sufijo alternante
+	diferencia de entropías
}	afijalidad
,	prueba de χ^2
;	razón de semejanza
.	coeficiente de Yule
:	información mutua

signo de algún índice no aparece en el vocablo de la columna de segmentaciones. quiere decir que el valor mayor de ese índice para ese vocablo fue de cero. La última columna contiene las veces en que acertaron uno o ambos de los índices de economía o entropía (se trata de, como se explicó arriba, los aciertos en una o ambas de las columnas ')' o '⟩').

Los criterios para determinar si las segmentaciones son correctas se aplicaron con flexibilidad, en especial con respecto al material que aparece entre las raíces y los sufijos. De esta manera, se dieron por buenas segmentaciones a ambos lados de las vocales temáticas. (quedando, por ejemplo, *~r* como única marca de infinitivo en 'abrasar', 'akonteser' y 'aplaudir'); de segmentos como *~g~* y *~k~* (en 'konbenga' y 'konduska'); de la vocal *~a~* en el sufijo *~al*, etc. Como se puede ver a lo largo de la misma tabla B.4 (véanse especialmente los núms. 682-718 y 426-438), este criterio favoreció en gran medida a las estadísticas de digramas (prueba de χ^2, razón de semejanza, coeficiente de Yule e información mutua), pero ni así obtuvieron tantos aciertos como los otros índices.

Finalmente y con respecto a las estadísticas de digramas, en los poquísimos vocablos de la muestra en que se observa el fenómeno de composición (véanse los núms. 842-845, al final de

la tabla B.4). estas cuatro estadísticas fueron los mejores índices de segmentación. El único compuesto para el que también los otros índices tuvieron éxito fue *beintiuna* (núm. 843). ya que lo segmentaron no entre las raíces involucradas. sino entre la base *beintiun~* y el sufijo *~a*. Son muy pocos vocablos compuestos para poder hablar con seguridad. pero esto apunta a que las fronteras entre las raíces o bases de un vocablo compuesto se pueden descubrir mejor con las estadísticas de digramas que con los índices para descubrir afijos. Valdría la pena aplicar y comparar estos índices en una muestra de vocablos compuestos.

Tabla B.4: Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:	
1	abarka	.a	a+bark⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
2	adkiera	.a	a+dki⟩er]}'),;.:a	1	1		1		1	1	1	1	1	1
3	agrada	.a	agr⟩;ad+]}'),.:a	1	1		1	1	1	1		1	1	1
4	akuosa	.a	a+kuos⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
5	alfombra	.a	a+lf]ombr⟩},;.:a			1			1	1	1	1	1	1
6	alkoba	.a	a+lko.b⟩]}'),;:a	1	1	1	1		1	1	1		1	1
7	almoada	.a	a+lmo⟩')ad]},;.:a	1					1	1	1	1	1	
8	amariya	.a	a+mariy⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
9	ansiana	.a	a+nsian⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
10	aplika	.a	aplik⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
11	aprenda	.a	a.;prend⟩+]}').:a	1	1	1	1	1	1			1	1	1
12	ariska	.a	a+ris.k⟩]}'),;:a	1	1	1	1		1	1	1		1	1
13	arkeolojika	.a	a+rkeolojik⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
14	asienta	.a	a+sient⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
15	bibora	.a	bibo]');.r⟩+},:a			1		1	1	1			1	1
16	blanda	.a	bl+an,;.:d⟩]}')a	1	1	1	1		1					1
17	brigada	.a	bri]g⟩+}')ad,;.:a							1	1	1	1	
18	dekanta	.a	de..:kan]}')t⟩+;a			1		1			1			1
19	desbentaja	.a	desbentaj⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
20	determina	.a	de;term+in⟩]}'),.:a	1	1	1	1		1	1		1	1	1
21	dibisa	.a	dibi,;s⟩+]}').:a	1	1	1	1	1	1			1	1	1
22	disfruta	.a	disfrut⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
23	dosena	.a	dos+en⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
24	dramatika	.a	dramatik⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
25	enerjetika	.a	en+erjetik⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
26	enjuaga	.a	en+juag⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
27	esfuersa	.a	e+sfuer,;s⟩]}').:a	1	1	1	1		1			1	1	1
28	eskueta	.a	eskue]}'),;.:t⟩+a			1		1						1
29	estupenda	.a	e+stupend⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
30	exacta	.a	e+xact⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
31	exkusa	.a	e+xkus⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
32	explosiba	.a	e+xplosib⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
33	extraordinaria	.a	extraordin+ari⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
34	gabardina	.a	gab,;.:ardin⟩+}a			1		1	1					1
35	gayeta	.a	gayet⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
36	ignominia	.a	ignomini⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
37	inbersa	.a	inbers⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
38	inbestiga	.a	in+besti,;g⟩]}').:a	1	1	1	1		1			1	1	1
39	inbita	.a	inbi,;t⟩+]}').:a	1	1	1	1	1	1			1	1	1
40	infraestructura	.a	infraestructur⟩+},;.:a			1		1	1	1	1	1	1	1
41	inkolora	.a	inkolor⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
42	inkomoda	.a	inkomod⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
43	intensa	.a	intens⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
44	jenerosa	.a	j+eneros⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
45	kalisa	.a	k.:a+li,;s⟩]}')a	1	1	1	1		1					1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:		
46	kantina	.a	ka+nti;.n⟩]}'),:a	1	1	1	1		1	1			1	1
47	kaptura	.a	kaptur⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
48	karabana	.a	kara]b'an⟩+},;.::a			1		1	1	1	1	1	1	1
49	karpeta	.a	ka+rp]et⟩}'),;.::a		1	1	1		1	1	1	1	1	1
50	kaserola	.a	ka+sero]}l⟩'),;.::a		1	1	1			1	1	1	1	1
51	klasika	.a	klasik⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
52	komplementaria	.a	komplementari⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
53	komplika	.a	kompl,;ik⟩+]}').:a	1	1	1	1	1	1			1	1	1
54	kondensa	.a	kondens⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
55	kontenta	.a	kontent⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
56	konyeba	.a	kon,;.:yeb⟩+}a			1		1	1					1
57	kulebra	.a	kulebr⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
58	lagrima	.a	lagri+m⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
59	laguna	.a	l+a,;.:gun⟩]}')a	1	1	1	1		1					1
60	lojika	.a	l,;.:oj+ik⟩]}')a	1	1	1	1		1					1
61	magnifika	.a	magnif+ik⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
62	maseta	.a	maset⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
63	mefenesina	.a	mefenesin⟩+},;.::a			1		1	1	1	1	1	1	1
64	misselanea	.a	mi+sselane⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
65	muneka	.a	munek⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
66	nocturna	.a	no+cturn⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
67	oaxakena	.a	o+axaken⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
68	obserba	.a	o+bser⟩b⟩]}',;.::a	1		1	1		1	1	1	1	1	1
69	ofensa	.a	ofen]s⟩+}'),;.::a		1	1	1	1	1	1	1	1	1	1
70	olimpika	.a	o+limpik⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
71	ontolojika	.a	o+ntolojik⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
72	oratoria	.a	oratori⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
73	pandiya	.a	pandiy⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
74	panoča	.a	pano,;.:č⟩+}a			1		1	1					1
75	papaya	.a	p+ap⟩ay]}'),;.::a	1	1		1		1	1	1	1	1	1
76	papila	.a	p+ap⟩i]}'),;.:la											
77	patolojika	.a	patolojik⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
78	patriotika	.a	pa+triotik⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
79	paulatina	.a	paulat+in⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
80	pediatra	.a	pe+diatr⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
81	penetra	.a	penetr⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
82	perfora	.a	pe+rfo,;r⟩]}').:a	1	1	1	1		1			1	1	1
83	plasenta	.a	plasent⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
84	plasma	.a	plas.m⟩+]}'),;::a	1	1	1	1	1	1	1	1		1	1
85	prebisora	.a	pre+bisor⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
86	presunta	.a	p+resunt⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1
87	primabera	.a	prima]'ver⟩+},;.::a			1		1	1	1	1	1	1	1
88	primitiba	.a	primitib⟩+]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
89	profusa	.a	p+rofu;.s⟩]}'),:a	1	1	1	1		1	1			1	1
90	proyecta	.a	proy+ect⟩]}'),;.::a	1	1	1	1		1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	:	.	:	
91	psikologa	.a	psikolog⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
92	sasona	.a	sason⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
93	sentroamerikana	.a	se+ntroamerikan⟩]}').:.:a	1	1	1	1		1	1	1	1	1	1
94	sibika	.a	sibik⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
95	siembra	.a	siemb+r⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
96	sifilitika	.a	si+filitik⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
97	siklika	.a	si+klik⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
98	simbolika	.a	si+mbolik⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
99	siniestra	.a	si+niestr⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
100	sinkronika	.a	si+nkronik⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
101	sintoma	.a	sinto]}')m⟩+,;.:a			1		1		1	1	1	1	1
102	sintura	.a	s,;.:intu+r⟩]}')a	1	1	1	1		1					1
103	siruela	.a	si+ruel⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
104	sobietika	.a	so+bietik⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
105	sosobra	.a	sos.ob+r⟩]}'),;:a	1	1	1	1		1	1	1		1	1
106	sospečosa	.a	so+spečos⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
107	tatema	.a	ta.;.:tem⟩+]}')a	1	1	1	1	1	1					1
108	tiniebla	.a	tiniebl⟩+},;.:a			1		1	1	1	1	1	1	1
109	transitoria	.a	tra+nsitori⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
110	tuerka	.a	tuer]}')k⟩+,;.:a			1		1		1	1	1	1	1
111	octaba	.ab.a	oct⟩ab+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
112	intensiba	.ib.a	intensi.b⟩+]}'),;:a	1	1	1	1	1	1	1	1		1	1
113	akumulatiba	.tib.a	a+kumulatib⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
114	aguada	.a.d.a	a+gu⟩a.:d]}'),;a	1	1	1	1		1	1	1	1	1	1
115	explikada	.a.d.a	explik⟩+a,;.:d]}')a	1	1	1	1	1	1	1	1	1	1	1
116	formada	.a.d.a	f+orm⟩a,;.:d]}')a	1	1	1	1		1	1	1	1	1	1
117	kreada	.a.d.a	kre⟩a,;d+]}').:a	1	1	1	1	1	1	1	1	1	1	1
118	plomada	.a.d.a	plom⟩]}',;.:a)d+a	1	1	1	1	1	1	1	1	1	1	1
119	poblada	.a.d.a	pobl⟩+a;d]}'),.:a	1	1	1	1	1	1	1	1	1	1	1
120	embestida	.i.d.a	embest+i)d⟩]}',;.:a	1	1	1	1	1	1	1	1	1	1	1
121	abansada	.ad.a	a+bans⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
122	akomodada	.ad.a	a+komod⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
123	akostada	.ad.a	a+kost⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
124	alejada	.ad.a	alej⟩+ad]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
125	almasenada	.ad.a	a+lmasen⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
126	alterada	.ad.a	a+lter⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
127	aplanada	.ad.a	aplan⟩+ad]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
128	apretada	.ad.a	a+pret⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
129	aproximada	.ad.a	a+proxim⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
130	aseptada	.ad.a	a+sept⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
131	asosiada	.ad.a	a+sosi⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
132	atropeyada	.ad.a	a+tropey⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
133	aumentada	.ad.a	a+ument⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1
134	belada	.ad.a	bel⟩ad+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
135	binkulada	.ad.a	bi+nkul⟩ad]}'),;.:a	1	1	1	1		1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	:	.	:	
136	deformada	.ad.a	de.::for+m⟩ad]}')a	1	1	1	1		1					1
137	deribada	.ad.a	derib⟩+ad]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
138	desbordada	.ad.a	desbord⟩+ad]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
139	destakada	.ad.a	destak⟩+ad]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
140	destinada	.ad.a	destin⟩+ad]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
141	deteriorada	.ad.a	deteri+or⟩ad]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
142	diborsiada	.ad.a	di+borsi⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
143	dokumentada	.ad.a	do+kument⟩ad]}'),;.::a	1	1	1	1		1	1	1	1	1	1
144	ebolusionada	.ad.a	e+bolusion⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
145	enkantada	.ad.a	e+nkant⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
146	enlatada	.ad.a	enlat⟩+ad]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
147	enmarkada	.ad.a	en.mark⟩+ad]}').::a	1	1	1	1	1	1	1	1		1	1
148	enmaskarada	.ad.a	enmaskar⟩+ad]}').;.::a	1	1	1	1	1	1	1	1	1	1	1
149	ensalada	.ad.a	ensal⟩+ad]}'),,.::a	1	1	1	1	1	1	1	1	1	1	1
150	entusiasmada	.ad.a	en+tusiasm⟩ad]}'),;.::a	1	1	1	1		1	1	1	1	1	1
151	espantada	.ad.a	e+spant⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
152	espesialisada	.ad.a	e+spesialis⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
153	estankada	.ad.a	e+stank⟩ad]}'),:.::a	1	1	1	1		1	1	1	1	1	1
154	estirada	.ad.a	e+st.:.::ir⟩ad]}')a	1	1	1	1		1					1
155	estudiada	.ad.a	estud+i⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
156	inkorporada	.ad.a	inkorp+or⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
157	instalada	.ad.a	in+stal⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
158	intensionada	.ad.a	in+tension⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
159	jirada	.ad.a	jir⟩ad+]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
160	kolokada	.ad.a	kolok⟩+ad]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
161	kondisionada	.ad.a	kondision⟩−.:.::ad]}')a	1	1	1	1	1	1	1	1	1	1	1
162	konectada	.ad.a	konec+t⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
163	labrada	.ad.a	la.:.::br⟩+ad]}')a	1	1	1	1	1	1					1
164	lansada	.ad.a	l.:ans⟩ad+]}'),,;a	1	1	1	1	1	1	1	1			1
165	morada	.ad.a	m.:.::or⟩ad+]}')a	1	1	1	1	1	1					1
166	pasada	.ad.a	pas⟩ad+]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
167	perfumada	.ad.a	perfum⟩ad+]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
168	pintada	.ad.a	pint⟩ad+]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
169	proyectada	.ad.a	proyect⟩ad+]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
170	sakrifikada	.ad.a	sa+krifik⟩ad]}').:.::a	1	1	1	1		1	1	1	1	1	1
171	salada	.ad.a	sal⟩.:.::ad+]}')a	1	1	1	1	1	1	1	1	1	1	1
172	solisitada	.ad.a	so+lisit⟩ad]}'),:.::a	1	1	1	1		1	1	1	1	1	1
173	superada	.ad.a	super⟩,,:.::ad+]}')a	1	1	1	1	1	1	1	1	1	1	1
174	tomada	.ad.a	tom⟩+ad]}').:.::a	1	1	1	1	1	1	1	1	1	1	1
175	tumbada	.ad.a	tumb⟩+ad]}'),:.::a	1	1	1	1	1	1	1	1	1	1	1
176	adkirida	.id.a	a+dkir⟩id]}'),:.::a	1	1	1	1		1	1	1	1	1	1
177	aludida	.id.a	a+ludid⟩]}'),,:.::a	1	1	1	1		1	1	1	1	1	1
178	debida	.id.a	de.:bid⟩+]}'),:a	1	1	1	1	1	1	1	1			1
179	desapersibida	.id.a	desapersib+id⟩]}'),;.::a	1	1	1	1	1	1	1	1	1	1	1
180	desidida	.id.a	desidid⟩+]}'),:.::a	1	1	1	1	1	1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:	
181	dibertida	.id.a	dibert+id⟩]}'),:.:a	1	1	1	1	1	1	1	1	1	1	1
182	difundida	.id.a	difundid⟩+]}'),:.:a	1	1	1	1	1	1	1	1	1	1	1
183	dirijida	.id.a	dirijid⟩+]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
184	finjida	.id.a	finjid⟩+]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
185	komplasida	.id.a	komplasid⟩+]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
186	komprometida	.id.a	komprometid⟩+]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
187	konstruida	.id.a	konstruid⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
188	kontraida	.id.a	kontra.:.:id⟩+]}')a	1	1	1	1	1	1	1	1	1	1	1
189	obstruida	.id.a	o+bstruid⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
190	ofresida	.id.a	o+fresid⟩]}').:.:a	1	1	1	1		1	1	1	1	1	1
191	sakudida	.id.a	sa+kudid⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
192	sernida	.id.a	se+r:.nid⟩]}').:a	1	1	1	1		1	1			1	1
193	bibienda	.iend.a	bi+biend⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
194	leyenda	.iend.a	le+yend⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
195	konbenga	.g.a	kon,;.:beng⟩+]}')a	1	1	1	1	1	1					1
196	kontenga	.g.a	kon,;.:ten]'g⟩+}a	1		1	1	1	1					1
197	konduska	.k.a	kondu})s]'k⟩+.;.:a	1		1	1	1		1	1	1	1	1
198	perteneska	.k.a	per+tenesk⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
199	astronomika	.ik.a	a+stronomik⟩]}').:.:a	1	1	1	1		1	1	1	1	1	1
200	atomika	.ik.a	a+tomik⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
201	biblika	.ik.a	bibl+ik⟩]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
202	botanika	.ik.a	botan]}i')k⟩+,;.:a			1		1		1	1	1	1	1
203	eroika	.ik.a	e+roik⟩]}').:.:a	1	1	1	1		1	1	1	1	1	1
204	filatelika	.ik.a	filatel+ik⟩]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
205	ironika	.ik.a	ironik⟩+]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
206	jenerika	.ik.a	jener.:ik⟩+]}'),;a	1	1	1	1	1	1	1	1	1	1	1
207	psikoanalitika	.ik.a	psikoanalit+ik⟩]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
208	dejenla	.en.l.a	de+jen,;.:l⟩]}')a	1	1	1	1		1	1	1	1	1	1
209	erirla	.i.r.l.a	e,;.:ri]}r')l⟩+a	1	1	1	1	1	1					1
210	kortarla	.ar.l.a	kort⟩ar.:l+]}'),;a	1	1	1	1	1	1	1	1	1	1	1
211	sakarla	.ar.l.a	sak⟩ar.:l+]}').:a	1	1	1	1	1	1	1	1	1	1	1
212	usarla	.ar.l.a	us⟩ar.;.:l+]}')a	1	1	1	1	1	1	1	1	1	1	1
213	kererla	.er.l.a	ke+rer.;.:l⟩]}')a	1	1	1	1		1	1	1	1	1	1
214	poderla	.er.l.a	p+oder.:l⟩]}'),;a	1	1	1	1		1	1	1	1	1	1
215	kumplirla	.ir.l.a	ku+mplir,;.:l⟩]}')a	1	1	1	1		1	1	1	1	1	1
216	arandela	.el.a	a+rand]}')el⟩,;.:a	1	1	1	1		1	1	1	1	1	1
217	bibisima	.isim.a	bi+bisim⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
218	guapisima	.isim.a	guapisim⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
219	primerisima	.isim.a	primer.isim⟩+]}'),;:a	1	1	1	1	1	1	1	1	1	1	1
220	italiana	.an.a	itali+an⟩]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
221	kristiana	.an.a	kristi+an⟩]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
222	brasileña	.eñ.a	brasil+eñ⟩]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
223	malageña	.eñ.a	malageñ⟩+},;.:a			1		1	1	1	1	1	1	1
224	ogareña	.eñ.a	o+gareñ⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
225	femenina	.in.a	femen+in⟩]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:	
226	salina	.in.a	sal,;i.:n⟩+]}')a	1	1	1	1	1	1	1	1			1
227	mantendrá	.dr.á	ma+n.;.:tendr]}')á	1	1		1		1					1
228	algodonera	.er.a	a+lgodoner⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
229	klabera	.er.a	klabe]}'),;.:r⟩+a			1		1						1
230	peskera	.er.a	p+esker⟩]}').:.:a	1	1	1	1		1	1	1	1	1	1
231	traisionera	.er.a	traisioner⟩+]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
232	salpikadera	.ad.er.a	salpikad]}').:.:er⟩+a	1	1	1	1	1	1	1	1	1	1	1
233	sonora	.or.a	son.:o+)r⟩]}'.:a	1		1	1		1	1	1	1	1	1
234	deboradora	.a.dor.a	de.bor+ador⟩]}'),;:a	1	1	1	1		1	1	1		1	1
235	konsiliadora	.a.dor.a	konsili+ador⟩]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
236	kriadora	.a.dor.a	kriado]}').:.:r⟩+a			1		1						1
237	moralisadora	.is.a.dor.a	moralis]}'a)dor⟩+.:.:a	1	1	1	1	1	1	1	1	1	1	1
238	potabilisadora	.is.a.dor.a	po,;.:tabilisador⟩+}a			1		1	1					1
239	edukadora	.ador.a	eduk+ador⟩]}'),;.:a	1	1	1	1	1	1	1	1	1	1	1
240	kalentura	.ur.a	ka+len')t]ur⟩},;.:a	1		1			1	1	1	1	1	1
241	bienesa	.es.a	bi+en⟩e,;s]}').:a	1	1	1	1		1			1	1	1
242	burgesa	.es.a	burg⟩+es]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
243	asombrosa	.os.a	a+sombros⟩]}'),;.:a	1	1	1	1		1	1	1	1	1	1
244	dudosa	.os.a	dudos⟩+]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
245	maldosa	.os.a	ma+ldos⟩]}').:.:a	1	1	1	1		1	1	1	1	1	1
246	peligrosa	.os.a	peligr+os⟩]}').;.:a	1	1	1	1	1	1	1	1	1	1	1
247	peresosa	.os.a	per+esos⟩]}').:.:a	1	1	1	1		1	1	1	1	1	1
248	motosikleta	.et.a	motosikl]}')et⟩+,;.:a	1	1	1	1	1	1	1	1	1	1	1
249	tarjeta	.et.a	tarj+et⟩]}').:.:a	1	1	1	1	1	1	1	1	1	1	1
250	embrita	.it.a	e,;.:mbr]}')it⟩+a	1	1	1	1	1	1					1
251	piedrita	.it.a	piedr]}').it⟩+.:.:a	1	1	1	1	1	1	1	1	1	1	1
252	pokita	.it.a	po,;k+it⟩]}').:a	1	1	1	1	1	1	1	1			1
253	puntita	.it.a	punt,;.:it⟩+]}')a	1	1	1	1	1	1	1	1	1	1	1
254	grandota	.ot.a	grand⟩o+t]}'),;.:a	1	1	1	1		1	1	1	1	1	1
255	muraya	.ay.a	mur⟩]a}')y+.:.:a	1		1		1		1	1	1	1	1
256	bentaniya	.iy.a	bentan]}')iy⟩+,;.:a	1	1	1	1	1	1	1	1	1	1	1
257	puntiya	.iy.a	p+untiy⟩]}').:.:a	1	1	1	1		1	1	1	1	1	1
258	abarkaba	.aba	abark⟩]}')a:.b+,:a	1	1	1	1		1					1
259	abrasaba	.aba	a+bras⟩]}')a:.b,:a	1	1	1	1		1					1
260	actuaba	.aba	actu⟩]}')a,:.:b+a	1	1	1	1		1					1
261	aguardaba	.aba	a,;.:guard⟩]}')ab+a	1	1	1	1		1					1
262	akonsejaba	.aba	a+konsej⟩]}')ab,;.:a	1	1	1	1		1					1
263	alejaba	.aba	al,;.:ej⟩]}')ab+a	1	1	1	1		1					1
264	alsaba	.aba	al,;.:s⟩]}')ab+a	1	1	1	1		1					1
265	aogaba	.aba	a,;.:og⟩]}')ab+a	1	1	1	1		1					1
266	aprobečaba	.aba	a+probeč⟩]}')ab,;.:a	1	1	1	1		1					1
267	asotaba	.aba	asot⟩]}')ab+,;.:a	1	1	1	1		1					1
268	bastaba	.aba	ba,;.:st⟩]}')ab+a	1	1	1	1		1					1
269	brotaba	.aba	bro.t⟩]}')ab+,;:a	1	1	1	1		1					1
270	dedikaba	.aba	dedik⟩]}')a:.b+,:a	1	1	1	1		1					1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:	
271	denotaba	.aba	de,;.::not⟩]}')ab+a	1	1	1	1		1					1
272	despertaba	.aba	despert⟩]}')a.b+.,::a	1	1	1	1		1					1
273	duraba	.aba	dur⟩]}'a,;.:b+)a	1		1	1		1					1
274	elebaba	.aba	el,;.:eb⟩]}')ab+a	1	1	1	1		1					1
275	emanaba	.aba	e,;::man⟩a]}').b+a			1								1
276	enbiaba	.aba	enbi⟩]}')a;.b+,:a	1	1	1	1		1					1
277	exportaba	.aba	export⟩+ab]}'),,:.:a			1		1						1
278	expulsaba	.aba	expuls⟩]}')a.b+,;:a	1	1	1	1		1					1
279	ignoraba	.aba	ignor⟩]}')a.b+,,::a	1	1	1	1		1					1
280	kantaba	.aba	k,;.:ant⟩]}')ab+a	1	1	1	1		1					1
281	kausaba	.aba	kaus⟩'}'),;a.:b+a	1	1	1	1		1	1	1			1
282	kebraba	.aba	ke,;.:br⟩]}')ab+a	1	1	1	1		1					1
283	kojeaba	.aba	ko,,::je⟩]}').ab+a	1	1	1	1		1			1		1
284	kompraba	.aba	kompr⟩]}'),,;.:ab+a	1	1	1	1		1	1	1	1	1	1
285	konsolaba	.aba	konsol⟩]}'),,:.:ab+a	1	1	1	1		1	1	1	1	1	1
286	kortaba	.aba	k,.:ort⟩}]');ab+a	1	1	1	1		1			1		1
287	markaba	.aba	mark⟩]}')a,;.:b+a	1	1	1	1		1					1
288	notaba	.aba	no.:t⟩]}'a,;b+)a	1		1	1		1					1
289	pagaba	.aba	pa.g⟩'}')ab+.::a	1	1	1	1		1					1
290	paraba	.aba	pa,,r⟩]}'a+.:b)a	1		1	1		1					1
291	sobraba	.aba	sobr⟩'}'),;.:ab+a	1	1	1	1		1	1	1	1	1	1
292	solisitaba	.aba	solisit⟩]}')a,:,b+:a	1	1	1	1		1					1
293	liturjia	.ia	li+turj]}')i⟩.:.:a	1	1		1		1					1
294	perfidia	.ia	perfi:.d]}').:i⟩+a	1	1		1		1	1			1	1
295	algarabía	.ía	a+lgara]}'),;.:bí⟩a											
296	balentía	.ent.ía	bal.:.:ent]}')í⟩+a	1	1		1		1	1	1	1	1	1
297	otelería	.ería	o,;::tel.e]}')rí⟩+a									1		
298	delikuensia	.sia	deli,;.:kuen]}')si⟩+a	1	1		1		1					1
299	inbariansia	.a.nsia	inbaria,;.:n]}')si⟩+a							1	1	1	1	
300	exijensia	.e.nsia	exij'e})nsi⟩+,;.:a	1	1		1		1					1
301	interferensia	.e.nsia	interfer]}')ensi⟩+,;.:a	1	1		1		1					1
302	koerensia	.e.nsia	k;o,.:er}')en]si⟩+a		1		1		1					1
303	konbibensia	.e.nsia	kon.bib]}e')nsi⟩+,,::a	1	1		1		1					1
304	permanensia	.e.nsia	permane,;.:n]}')si⟩+a							1	1	1	1	
305	preferensia	.e.nsia	prefer]}'en)si⟩+,;.:a	1			1		1					
306	ekibalensia	.ensia	ekibal]}')ensi⟩+,;.:a	1	1		1		1					1
307	aumentara	.a.ra	a+ument⟩]},;ar').:a	1		1			1	1	1			1
308	explikara	.a.ra	explik⟩+'),;ar]}.:a		1	1	1	1		1	1			1
309	lebantara	.a.ra	le+bant⟩'),;ar]}.:a		1	1	1			1	1			1
310	produjera	.e.ra	produje]}')r⟩+,;.:a	1	1		1		1					1
311	leyera	.ie.ra	ley+e.:r⟩]}'),;a					1				1	1	
312	tubiera	.ie.ra	tubier⟩+]}'),;.:a											
313	delikadesa	.esa	delikad⟩]}').es+.,::a	1	1	1	1		1			1		1
314	tasita	.ita	ta,;.:s]}')it⟩+a	1	1	1	1	1	1					1
315	nobenta	.nta	no.ben]}')t⟩+,;::a											

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:	
316	umanista	.ista	uman]}ist⟩+'),;.:a	1					1					
317	kansionista	.ista	kansion]}')ist⟩+,;.:a	1	1		1		1					1
318	susiedad	.edad	s+u,;.:si⟩]}')edad	1	1	1	1		1					1
319	deformidad	.idad	d+e,.:form⟩]}');idad	1	1	1	1		1		1			1
320	disparidad	.idad	di+spar⟩]}'),;.:idad	1	1	1	1		1	1	1	1	1	1
321	familiaridad	.idad	fa+miliar⟩]}'),;.:idad	1	1	1	1		1	1	1	1	1	1
322	inkapasidad	.idad	in+,;.:kapas⟩]}')idad	1	1	1	1		1					1
323	probidad	.idad	pro+,;b⟩]}'.:idad	1		1	1		1			1	1	1
324	dualidad	.al.idad	d).:u⟩,;al+]}'idad	1		1	1	1	1	1	1			1
325	espiritualidad	.al.idad	e+spiritu⟩al]}'),;.:idad	1	1	1	1		1	1	1	1	1	1
326	unibersalidad	.al.idad	u+nibers⟩al]}'),;.:idad	1	1	1	1		1	1	1	1	1	1
327	morbilidad	.il.idad	mo+r,;.:b]'il⟩}idad	1		1	1		1					1
328	solubilidad	.bil.idad	so+lu]}',;.:bil⟩i)dad	1		1	1		1	1	1	1	1	1
329	electrisidad	.is.idad	e+lectr⟩i];.s}')i,:dad		1	1	1		1					1
330	akomode	.e	a+komo,;d⟩]}').:e	1	1	1	1		1			1	1	1
331	apriete	.e	a+pri⟩et]}'),;.:e	1	1		1		1	1	1	1	1	1
332	aprobeče	.e	a+probe;č⟩]}'),.:e	1	1	1	1		1	1		1	1	1
333	bertise	.e	b+ert⟩i]}'),;.:se											
334	dibierte	.e	dibi⟩er+t]}'),;.:e	1	1		1		1	1	1	1	1	1
335	ejekute	.e	e+jeku),;t⟩]}'.:e	1		1	1		1			1	1	1
336	enkaje	.e	en+k⟩aj]}'),;.:e	1	1		1		1	1	1	1	1	1
337	enmudese	.e	enmud⟩+es]}'),;.:e	1	1		1		1	1	1	1	1	1
338	galope	.e	ga+lo⟩p]}'),;.:e	1	1		1		1	1	1	1	1	1
339	inbalide	.e	in+balid⟩]}'),;.:e	1	1	1	1		1	1	1	1	1	1
340	induse	.e	in.;dus⟩+]}').:e	1	1	1	1	1	1			1	1	1
341	inerme	.e	in⟩,;.:er]}')m+e					1						
342	inisie	.e	in,;is⟩+i]}').:e	1	1		1		1			1	1	1
343	inmueble	.e	inmuebl⟩+},;.:e			1		1	1	1	1	1	1	1
344	inside	.e	i+nsi.:d⟩]}'),;e	1	1	1	1		1	1	1			1
345	jarabe	.e	jar⟩a]}'b⟩+,;.:e			1		1		1	1	1	1	1
346	kakauate	.e	kakau⟩+}at,;.:e							1	1	1	1	
347	katastrofe	.e	ka+tastrof⟩]}'),;.:e	1	1	1	1		1	1	1	1	1	1
348	kombate	.e	ko+mb⟩a')t]},;.:e	1					1	1	1	1	1	
349	konduse	.e	k+ondus⟩]}',;.:e	1		1	1		1	1	1	1	1	1
350	konssiente	.e	kons]}')s⟩+ient,;.:e							1	1	1	1	
351	padese	.e	pad⟩+)es]}',;.:e	1			1		1	1	1	1	1	
352	patente	.e	pat,;.:en+t⟩]}')e	1	1	1	1		1					1
353	prebalese	.e	prebal⟩+es]}'),;.:e	1	1		1		1	1	1	1	1	1
354	proporsione	.e	pro+porsio,;n⟩]}').:e	1	1	1	1		1			1	1	1
355	silise	.e	sil⟩is+]}'),;.:e	1	1		1	1	1	1	1	1	1	1
356	sobrebiene	.e	so+bre,;.:bi]'en⟩}e			1			1					1
357	suspende	.e	su+spen,;d⟩]}').:e	1	1	1	1		1			1	1	1
358	tepače	.e	te,;.:p⟩]}'ač+e					1						
359	timbre	.e	ti,;.:mb⟩+r]}')e	1	1		1		1					1
360	trassiende	.e	trassiend⟩+},;.:e			1		1	1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	:	.	:	
361	triple	.e	t+rip]}')l⟩,;.:e			1				1	1	1	1	1
362	unanime	.e	u+nanim)]}'),;.:e	1	1	1	1		1	1	1	1	1	1
363	kitandole	.a.ndo.l.e	kit⟩+andol]}'),;.:e	1	1	1	1	1	1	1	1	1	1	1
364	ganarle	.ar.l.e	g+an⟩ar;.l]}'),.:e	1	1	1	1		1	1	1	1	1	1
365	mostrarle	.ar.l.e	mo+str⟩ar.l]}'),;:e	1	1	1	1		1	1	1	1	1	1
366	kapote	.ot.e	k.;.:a+p)]ot}')e	1	1	1	1		1					1
367	fuselaje	.aje	fu.;.:sel⟩+}aje			1		1	1					1
368	enbidiable	.a.ble	en+bidi)]}')abl,;.:e	1	1	1	1		1					1
369	indeseable	.a.ble	inde+se⟩a]}')bl,;.:e	1	1	1	1		1	1	1	1	1	1
370	inexorable	.a.ble	inexor⟩+}abl,;.:e			1		1	1					1
371	konsiderable	.a.ble	konsider)]}ab')l,;.:e	1		1			1					1
372	prebisible	.i.ble	prebis)]}'ibl+,;.:e	1		1	1		1					1
373	desfaborable	.able	desfabor⟩+}abl,;.:e			1		1	1					1
374	impekable	.able	impe,;.:k⟩+}able			1		1	1					1
375	sakame	.a.me	sak⟩+a].:m}'),;e	1		1		1				1	1	1
376	sueltame	.a.me	su+elt⟩a]}'),;.:me	1	1	1	1		1	1	1	1	1	1
377	detenerme	.e.r.me	de.;.:ten⟩e')r]}m+e	1	1	1	1		1					1
378	sentirme	.i.r.me	se+nt)i]}')r,;.:me	1	1	1	1		1	1	1	1	1	1
379	dejarme	.ar.me	dej)]}')ar,;.:m+e	1	1	1	1		1	1	1	1	1	1
380	ganarme	.ar.me	gan)]}')ar,;.:m+e	1	1	1	1		1	1	1	1	1	1
381	pudiese	.ie.se	pudi⟩e]}'),;.:s+e	1	1		1		1	1	1	1	1	1
382	burlarse	.a.r.se	bu+rl)]}'a.r),;:se	1	1	1	1		1	1	1	1	1	1
383	bolberse	.e.r.se	bo+lb⟩e')r]},;.:se	1	1	1	1		1	1	1	1	1	1
384	imponerse	.e.r.se	impon⟩e]}')r:.s+,:e	1	1	1	1		1		1	1		1
385	perderse	.e.r.se	perd⟩e')r]},:.:s+e	1	1	1	1		1	1	1	1	1	1
386	dormirse	.i.r.se	do+rm⟩i]}')r,;.:se	1	1	1	1		1	1	1	1	1	1
387	eskribirse	.i.r.se	eskrib⟩i')r]},;.:s+e	1	1	1	1		1	1	1	1	1	1
388	gastarse	.ar.se	gast⟩]}')ar,;.:s+e	1	1	1	1		1	1	1	1	1	1
389	kasarse	.ar.se	kas⟩]},;.:ar')s+e	1	1	1	1		1	1	1	1	1	1
390	komunikarse	.ar.se	komunik⟩]}')ar.s+,;:e	1	1	1	1		1			1		1
391	prestarse	.ar.se	pr⟩est⟩]}'ar,;.:s+e	1		1	1		1	1	1	1	1	1
392	sakarse	.ar.se	sak⟩]}')ar,;.:s+e	1	1	1	1		1	1	1	1	1	1
393	tratarse	.ar.se	trat⟩+]}')ar,;.:se	1	1	1	1	1	1	1	1	1	1	1
394	alusinante	.a.nte	a+lusin⟩a]}')nt,;.:e	1	1	1	1		1					1
395	apasionante	.a.nte	a+pasion⟩a.n]}')t,;:e			1						1		1
396	extrabagante	.a.nte	extra+bag⟩}an]'t,;.:e			1			1					1
397	fabrikante	.a.nte	fabrik⟩an]}')t+,;.:e			1								1
398	fassinante	.a.nte	fa+ssin⟩a.n]}')t,;:e			1						1		1
399	flotante	.a.nte	flot⟩a.n]}')t+,;:e			1						1		1
400	oxidante	.a.nte	o+xid⟩a.n]}')t,;:e			1						1		1
401	pedante	.a.nte	ped)]}'a,;.:nt+e	1		1	1		1	1	1	1	1	1
402	protestante	.a.nte	pro+test⟩an]}')t,;.:e			1								1
403	imponente	.e.nte	impone];.n}')t⟩+,:e	1							1	1		
404	inkandessente	.e.nte	in+kandessen]}')t⟩,;.:e											
405	kompetente	.e.nte	kompete')n]};.t⟩+,:e		1		1							1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	.	;	.	:	
406	bibiente	.ie.nte	bib⟩]i.:.:en}')t+e	1		1								1
407	saliente	.ie.nte	s.:al⟩.:ie]}n')t+e	1		1			1	1	1			1
408	sobresaliente	.ie.nte	so+bresal⟩]ie')n}t,;.:e	1	1	1	1							1
409	fielmente	.mente	fie,:.:l]}')ment⟩+e	1	1		1		1					1
410	ostensiblemente	.mente	o+stensible]}').:.:ment⟩e	1	1		1		1	1	1	1	1	1
411	enteramente	.a.mente	e+nt,;.:er⟩]}'am)ente	1		1	1		1					1
412	injenuamente	.a.mente	in+jenu⟩]}'),;.:amente	1	1	1	1		1	1	1	1	1	1
413	konkretamente	.a.mente	konkret⟩]}').ament+,;:e	1	1	1	1		1			1		1
414	magnifikamente	.a.mente	ma+gnifik⟩]]}').:.:amente	1	1	1	1		1	1	1	1	1	1
415	practikamente	.a.mente	pract+ik⟩]}')a.:.:mente	1	1	1	1		1	1	1	1	1	1
416	selosamente	.a.mente	s,:e:.los⟩]}a')ment+e	1	1	1	1		1					1
417	sensiyamente	.a.mente	se+nsiy⟩]},:.:a')mente	1	1	1	1		1	1	1	1	1	1
418	silensiosamente	.a.mente	si+lensios⟩]}',:.:a)mente	1	1	1	1		1	1	1	1	1	1
419	tiernamente	.a.mente	tiern⟩]}',;.:a)ment+e	1	1	1	1		1	1	1	1	1	1
420	koordinadamente	.ad.a.mente	koordin⟩ad]},:.:a')mente	1	1	1	1		1	1	1	1	1	1
421	detenidamente	.id.a.mente	deten⟩id]}'),:.:ament+e	1	1	1	1		1	1	1	1	1	1
422	perdidamente	.id.a.mente	perd⟩id]},:.:a')ment+e	1	1	1	1		1	1	1	1	1	1
423	malisiosamente	.os.a.mente	malisios⟩]}')a.:.:ment+e	1	1	1	1		1	1	1	1	1	1
424	desfaborablemente	.able.mente	desfabor⟩+able]}'),;.:mente	1	1	1	1	1	1	1	1	1	1	1
425	bolbiste	.i.ste	bol+b⟩]i}'),;.:ste	1	1	1	1		1	1	1	1	1	1
426	arterial	.a.l	a+rteri⟩a]}'),:.:l	1	1	1	1		1	1	1	1	1	1
427	dominikal	.a.l	do+minik⟩]]}')a.:.:l	1	1	1	1		1	1	1	1	1	1
428	ejidal	.a.l	e+jid⟩]]}')a,:.:l	1	1	1	1		1	1	1	1	1	1
429	experimental	.a.l	e+xperiment⟩]}')a,;.:l	1	1	1	1		1	1	1	1	1	1
430	flubial	.a.l	flubi⟩+}a,:.:l			1		1	1	1	1	1	1	1
431	karnal	.a.l	ka+rn⟩]}',;.:a)l	1	1	1	1		1	1	1	1	1	1
432	konjuntibal	.a.l	kon+juntib⟩]}'a).;.:l	1	1	1	1		1	1	1	1	1	1
433	koronal	.a.l	koron⟩+}')a].:.:l	1	1	1	1	1	1	1	1	1	1	1
434	nupsial	.a.l	nupsi⟩+a]}'),;.:l	1	1	1	1	1	1	1	1	1	1	1
435	pastoral	.a.l	pa+stor⟩]},;.:a')l	1	1	1	1		1	1	1	1	1	1
436	selestial	.i.a.l	se+lest]}i)')a,:.:l	1	1	1	1		1	1	1	1	1	1
437	sensorial	.i.a.l	se+nsori⟩]}')a.:.:l	1	1	1	1		1	1	1	1	1	1
438	substansial	.i.a.l	su+bstansi⟩a]}'),;.:l	1	1	1	1		1	1	1	1	1	1
439	femenil	.i.l	femen⟩+i]}'),;.:l	1	1	1	1	1	1	1	1	1	1	1
440	bentral	.al	b,;.:e+nt]'r⟩}al			1			1					1
441	estatal	.al	e+sta,;.:t⟩]}'al	1		1	1		1					1
442	kaporal	.al	ka,;.:p]'or⟩+}al			1		1	1					1
443	kordial	.al	k,;.:o+rd]'i⟩}al			1			1					1
444	matrimonial	.al	ma+trimoni⟩]}'),;.:al	1	1	1	1		1	1	1	1	1	1
445	multilateral	.al	mu+lt]'ilater⟩}a,;.:l			1			1					1
446	multinasional	.al	mu+ltinasion⟩}a,;.:l			1			1					1
447	trassendental	.al	tra+ssendent⟩]}').a,;:l	1	1	1	1		1			1		1
448	abitasional	.sion.al	a+,:.:bitasion⟩]}')al	1	1	1	1		1					1
449	klabel	.el	klab⟩+,:.:e]}')l			1		1		1	1	1	1	1
450	afrontan	.a.n	a+front⟩')a]}.:.:n	1	1	1	1		1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	~	}	,	;	.	:	
451	dependan	.a.n	depend⟩]}'),;.:a+n	1	1	1	1	1	1	1	1	1	1	1
452	desidan	.a.n	de+sid⟩]},:.:a')n	1	1	1	1		1	1	1	1	1	1
453	designan	.a.n	des+ign⟩]}')a,;.:n	1	1	1	1		1	1	1	1	1	1
454	enfrentan	.a.n	en+frent⟩]}')a.:.:n	1	1	1	1		1	1	1	1	1	1
455	enkantan	.a.n	e+nkant⟩]}')a.:.:n	1	1	1	1		1	1	1	1	1	1
456	espesifikan	.a.n	e+spesifik⟩]}'):a..:n	1	1	1	1		1	1	1	1	1	1
457	estiman	.a.n	estim⟩+]}')a,;.:n	1	1	1	1	1	1	1	1	1	1	1
458	garantisan	.a.n	ga+rantis⟩]}')a.:.:n	1	1	1	1		1	1	1	1	1	1
459	guardan	.a.n	g+uard⟩]}')a,;.:n	1	1	1	1		1	1	1	1	1	1
460	kieran	.a.n	k),;.:i+er⟩]}a'n	1		1	1		1					1
461	konkretan	.a.n	konkre+t⟩}'),.:a].:n	1	1	1	1		1	1	1	1	1	1
462	piensan	.a.n	pien+s⟩]}'),;.:an	1	1	1	1		1	1	1	1	1	1
463	preokupan	.a.n	preoku+p⟩]}')a.:.:n	1	1	1	1		1	1	1	1	1	1
464	priban	.a.n	pr,;ib⟩+}')a].:n	1	1	1	1	1	1			1	1	1
465	soportan	.a.n	s+oport⟩]}')a,;.:n	1	1	1	1		1	1	1	1	1	1
466	trasan	.a.n	tr+)a,.:s⟩]}'a.:n	1		1	1		1			1	1	1
467	estayaban	.aba.n	est.;.:ay⟩]}')ab+an	1	1	1	1		1					1
468	kejaban	.aba.n	k.:e,;j⟩}ab+a]')n	1	1	1	1		1					1
469	sentaban	.aba.n	se)+nt⟩.:.:ab}a]'n	1		1	1			1	1	1	1	1
470	ordenaran	.a.ra.n	o+rden⟩]}a,;.:r')an	1		1			1	1	1	1	1	1
471	bolbieran	.ie.ra.n	bo+lbier⟩a]}'),.:.:n	1	1	1	1		1	1	1	1	1	1
472	kisieran	.ie.ra.n	kisier⟩+a]}'),;.:n	1	1		1		1	1	1	1	1	1
473	kreyeran	.ie.ra.n	kreyer⟩+a]}').:.:n	1	1	1	1	1	1	1	1	1	1	1
474	enteren	.e.n	e+nt,;er⟩]}.:e')n	1	1	1	1		1			1	1	1
475	expliken	.e.n	explik⟩+]}.:e').:n	1	1	1	1	1	1	1	1	1	1	1
476	extraen	.e.n	e+xtra⟩.:e]}').:n	1	1	1	1		1	1	1	1	1	1
477	formen	.e.n	form⟩+j}..:e'):n	1	1	1	1	1	1	1	1	1	1	1
478	funsionen	.e.n	funsio,:n⟩+]}.:e')n	1	1	1	1	1	1			1	1	1
479	komponen	.e.n	k+ompon⟩e]}').:.:n	1	1	1	1		1	1	1	1	1	1
480	presenten	.e.n	pre+sent⟩]}.:e').:n	1	1	1	1		1	1	1	1	1	1
481	proiben	.e.n	p+roib⟩},;.:e]')n	1	1	1	1		1	1	1	1	1	1
482	suponen	.e.n	s+upon⟩.:e]}').:n	1	1	1	1		1	1	1	1	1	1
483	surten	.e.n	s+u,;.:rt⟩e]}')n	1	1	1	1		1					1
484	suspenden	.e.n	su+spend⟩,;.:e]}')n	1	1	1	1		1	1	1	1	1	1
485	intersesión	.ión	in+ter⟩se]',;.:s})ión		1				1					1
486	aflicsión	.sión	a+fli.c]}'),;:si⟩ón	1	1		1		1	1	1		1	1
487	disecsión	.sión	dis⟩e,;.:c]}')sió+n	1	1		1		1					1
488	konbecsión	.sión	kon+b⟩e,;.:c]}')sión	1	1		1		1					1
489	intubasión	.a.sión	int,;.:u]'b⟩+}asión			1		1	1					1
490	ignisión	.i.sión	ign⟩]}'),;.:isió+n	1	1	1	1		1	1	1	1	1	1
491	alkansaron	.a.ron	a+lkans⟩]}ar,;.:o')n	1		1			1					1
492	ayaron	.a.ron	ay⟩]}'),;.:aro+n	1	1	1	1		1	1	1	1	1	1
493	bajaron	.a.ron	baj⟩]}').:.:aro+n	1	1	1	1		1	1	1	1	1	1
494	destakaron	.a.ron	destak⟩+]}')a.:.:ron	1	1	1	1	1	1	1	1	1	1	1
495	estimaron	.a.ron	e:stim⟩]}'),;.aro+n	1	1	1	1		1	1	1	1		1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	:	.	:	
496	estudiaron	.a.ron	e+studi)]}')ar,;.:on	1	1	1	1		1					1
497	gritaron	.a.ron	grit⟩]}').ar.::o+n	1	1	1	1		1					1
498	inkontraron	.a.ron	i,;.:n+kontr⟩ar]}')on			1								1
499	kantaron	.a.ron	k,;.:ant⟩ar+o]}')n			1								1
500	klabaron	.a.ron	k,;.:lab)]}').aro+n	1	1	1	1		1			1		1
501	komensaron	.a.ron	komens⟩+]}')ar,;.:on	1	1	1	1	1	1					1
502	komprobaron	.a.ron	komprob⟩+]}')ar,;.:on	1	1	1	1	1	1					1
503	kontinuaron	.a.ron	kon+tinu)]}')ar,;.:on	1	1	1	1		1					1
504	krusaron	.a.ron	k,:rus)]}')ar;.o+n	1	1	1	1		1					1
505	mataron	.a.ron	ma,;.:t)]}')ar+on	1	1	1	1		1					1
506	proporsionaron	.a.ron	pro+porsion)]}')ar,;.:on	1	1	1	1		1					1
507	tokaron	.a.ron	tok⟩+]}')ar,;.:on	1	1	1	1	1	1					1
508	trasladaron	.a.ron	tras+lad⟩]}')ar,;.:on	1	1	1	1		1					1
509	aborto	.o	a+bort⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
510	aerodromo	.o	aer]'odrom⟩},,;.:o			1			1	1	1	1	1	1
511	aeropuerto	.o	a+eropuert⟩},,;.:o			1			1	1	1	1	1	1
512	alimento	.o	a+limen)t⟩]}',,;.:o	1		1	1		1	1	1	1	1	1
513	anunsio	.o	an+unsi)]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
514	arsobispo	.o	a+rsobisp)]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
515	asesino	.o	a+ses⟩in]}'),,;.:o	1	1		1		1	1	1	1	1	1
516	barlobento	.o	ba+rlobent⟩},,;.:o			1			1	1	1	1	1	1
517	bastardo	.o	bastar]}'),,;.:d⟩o			1								1
518	bentrikulo	.o	be+ntrikul)]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
519	besino	.o	bes⟩in+]}'),,;.:o	1	1		1	1	1	1	1	1	1	1
520	buelto	.o	bu+elt⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
521	deskontento	.o	deskont]'ent⟩+},,;.:o			1		1	1	1	1	1	1	1
522	despido	.o	desp⟩,,;.:id+]}')o	1	1		1	1	1					1
523	duodeno	.o	duod⟩+en]}'),,;.:o	1	1		1		1	1	1	1	1	1
524	ebento	.o	eben]}')t⟩+.;.:o			1		1		1	1	1	1	1
525	filtro	.o	f+iltr⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
526	fisiko	.o	fisik⟩+]}'),,;.:o	1	1	1	1	1	1	1	1	1	1	1
527	inberso	.o	inber,;s⟩+]}').:o	1	1	1	1	1	1			1	1	1
528	injenuo	.o	in+jenu)]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
529	inkomodo	.o	ink+omod⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
530	inospito	.o	inosp⟩+it]}'),,;.:o	1	1		1		1	1	1	1	1	1
531	insiso	.o	ins⟩is+]}'),,;.:o	1	1		1	1	1	1	1	1	1	1
532	konsekutibo	.o	konse+kutib⟩]}'),,;.:o	1	1	1	1	1	1	1	1	1	1	1
533	kritiko	.o	krit+ik⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
534	magnetiko	.o	ma+gnetik⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
535	murmuyo	.o	mu+rmu]'y⟩},,;.:o			1			1	1	1	1	1	1
536	negatibo	.o	n+egatib⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
537	pabimento	.o	pa+biment⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
538	parametro	.o	pa+rametr⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
539	polisiako	.o	poli+siak⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1
540	postibo	.o	p+ostib⟩]}'),,;.:o	1	1	1	1		1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:	
541	sakramento	.o	sa+krament⟩]}'),;.:o	1	1	1	1		1	1	1	1	1	1
542	seudonimo	.o	se+udonim⟩},:.:o			1			1	1	1	1	1	1
543	simbolo	.o	simbol⟩+]}'),;.:o	1	1	1	1	1	1	1	1	1	1	1
544	sinnumero	.o	sin,:.:num⟩+}ero											
545	sintetiko	.o	si+ntetik⟩]}').:.:o	1	1	1	1		1	1	1	1	1	1
546	sokoro	.o	s+okor⟩]}'),:.:o	1	1	1	1		1	1	1	1	1	1
547	sotano	.o	s+o.t⟩an]}'),;:o	1	1		1		1	1	1		1	1
548	telegrafo	.o	te+legraf⟩]}'),:.:o	1	1	1	1		1	1	1	1	1	1
549	tirano	.o	tir⟩+a.;n]}').:o	1	1		1		1			1	1	1
550	trienio	.o	t,;:ri.eni⟩+}o			1		1	1					1
551	ampliado	.a.d.o	a+mpli⟩,;a.:d]}')o	1	1	1	1		1	1	1	1	1	1
552	anotado	.a.d.o	anot⟩+a;.d]}'),:o	1	1	1	1	1	1	1	1	1	1	1
553	apoderado	.a.d.o	a+poder⟩]}')a;.d,:o	1	1	1	1		1	1	1	1	1	1
554	apoyado	.a.d.o	a+poy⟩a.d]}'),;:o	1	1	1	1		1	1	1	1	1	1
555	ayudado	.a.d.o	a+yud⟩]}a,;.:d')o	1	1	1	1		1	1	1	1	1	1
556	estimulado	.a.d.o	estimu+l⟩a,;.:d]}')o	1	1	1	1		1	1	1	1	1	1
557	exkusado	.a.d.o	exkus⟩]}')a,;.:d+o	1	1	1	1	1	1	1	1	1	1	1
558	finado	.a.d.o	f⟩i.;n⟩]}a'.:d+o	1		1	1	1	1			1	1	1
559	kunado	.a.d.o	k.;u.:n⟩a]}')d+o	1	1	1	1	1	1					1
560	liberado	.a.d.o	li+ber⟩a,;.:d]}')o	1	1	1	1		1	1	1	1	1	1
561	perfecsionado	.a.d.o	perfecsion⟩]}')a;.d+.:o	1	1	1	1	1	1	1	1	1	1	1
562	platikado	.a.d.o	p+latik⟩]}'a,;.:d)o	1	1	1	1		1	1	1	1	1	1
563	presentado	.a.d.o	present⟩a,;.:d+]}')o	1	1	1	1	1	1	1	1	1	1	1
564	solisitado	.a.d.o	so+lisit⟩a,;.d]}'):o	1	1	1	1		1	1	1	1	1	1
565	desaparesido	.i.d.o	des+apares⟩i.d]}'),;:o	1	1	1	1		1	1	1	1	1	1
566	fluido	.i.d.o	flu⟩i)d+]}',;.:o	1	1	1	1	1	1	1	1	1	1	1
567	insistido	.i.d.o	insist⟩i]}'),;.:d+o	1	1	1	1	1	1	1	1	1	1	1
568	konkluido	.i.d.o	konkl+u⟩i.d]}'),;:o	1	1	1	1		1	1	1	1	1	1
569	konsegido	.i.d.o	konseg⟩+i],;.d}'):o	1	1	1	1	1	1	1	1	1	1	1
570	serbido	.i.d.o	se+rb⟩i],;.:d}')o	1	1	1	1		1	1	1	1	1	1
571	surjido	.i.d.o	surj⟩+i.d]}'),;:o	1	1	1	1	1	1	1	1	1	1	1
572	adekuado	.ad.o	a+deku⟩ad]}'),;.:o	1	1	1	1		1	1	1	1	1	1
573	adornado	.ad.o	a+dorn⟩ad]}'),;.:o	1	1	1	1		1	1	1	1	1	1
574	ajitado	.ad.o	a+jit⟩ad]}'),;.:o	1	1	1	1		1	1	1	1	1	1
575	asombrado	.ad.o	a+sombr⟩ad]}'),;.:o	1	1	1	1		1	1	1	1	1	1
576	autorisado	.ad.o	a+utoris⟩ad]}'),;.:o	1	1	1	1		1	1	1	1	1	1
577	bolado	.ad.o	bo,;l⟩.:ad+]}')o	1	1	1	1	1	1			1	1	1
578	bordado	.ad.o	bord⟩+ad]}'),;.:o	1	1	1	1	1	1	1	1	1	1	1
579	desgastado	.ad.o	desgast⟩]}')ad+,;.:o	1	1	1	1	1	1	1	1	1	1	1
580	despedasado	.ad.o	despedas⟩+ad]}'),;.:o	1	1	1	1	1	1	1	1	1	1	1
581	despreokupado	.ad.o	despreokup⟩+ad]}'),;.:o	1	1	1	1	1	1	1	1	1	1	1
582	disgustado	.ad.o	di+sgust⟩ad]}'),;.:o	1	1	1	1		1	1	1	1	1	1
583	drenado	.ad.o	d.;.:ren⟩]}')ad+o	1	1	1	1	1	1					1
584	entrado	.ad.o	entr⟩+.;.:ad]}')o	1	1	1	1	1	1	1	1	1	1	1
585	inklinado	.ad.o	inklin⟩+ad]}').;.:o	1	1	1	1	1	1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:	
586	kolejiado	.ad.o	koleji⟩]}').:.:ad+o	1	1	1	1	1	1	1	1	1	1	1
587	konsertado	.ad.o	k,;.:onsert⟩+ad]}')o	1	1	1	1	1	1					1
588	perfilado	.ad.o	perfil⟩]}').:.:ad+o	1	1	1	1	1	1	1	1	1	1	1
589	soldado	.ad.o	sold⟩+ad]}').:.:o	1	1	1	1	1	1	1	1	1	1	1
590	titulado	.ad.o	titul⟩+ad]}').:.:o	1	1	1	1	1	1	1	1	1	1	1
591	tostado	.ad.o	to,;.:st⟩+ad]}')o	1	1	1	1	1	1					1
592	bensido	.id.o	b+en;.s⟩id]}').:.:o	1	1	1	1		1	1			1	1
593	enduresido	.id.o	en+dures⟩id]}').:.:o	1	1	1	1		1	1	1	1	1	1
594	ensendido	.id.o	e+nsend⟩id]}'),:.:o	1	1	1	1	1	1	1	1	1	1	1
595	entretenido	.id.o	en+tre,;.:ten⟩id]}')o	1	1	1	1		1					1
596	komprimido	.id.o	komprim⟩+id]}').:.:o	1	1	1	1	1	1	1	1	1	1	1
597	meresido	.id.o	me,;.:res⟩id+]}')o	1	1	1	1	1	1					1
598	mobido	.id.o	mob⟩+i,;.:d]}')o	1	1	1	1	1	1	1	1	1	1	1
599	pertenesido	.id.o	per+tenes⟩]}'),;.:ido	1	1	1	1		1	1	1	1	1	1
600	tejido	.id.o	tej⟩id+]}').:.:o	1	1	1	1	1	1	1	1	1	1	1
601	imajinario	.ari.o	imajinari⟩+]}'),:.:o	1	1	1	1	1	1	1	1	1	1	1
602	intermediario	.ari.o	inter+medi]}')ari),;.:o	1	1	1	1		1	1	1	1	1	1
603	kalbario	.ari.o	ka+lb]}').:.:ari⟩o	1	1	1	1		1	1	1	1	1	1
604	tersiario	.ari.o	te+rsi]}'ar⟩i⟩.:.:o	1		1	1		1	1	1	1	1	1
605	arkaiko	.ik.o	a+rka]}')ik⟩.:.:o	1	1	1	1		1	1	1	1	1	1
606	artistiko	.ik.o	a+rtistik⟩]}'),;.:o	1	1	1	1		1	1	1	1	1	1
607	demokratiko	.ik.o	demokrat+ik⟩]}'),:.:o	1	1	1	1	1	1	1	1	1	1	1
608	idrauliko	.ik.o	idraul+ik⟩]}'),;.:o	1	1	1	1	1	1	1	1	1	1	1
609	ironiko	.ik.o	ironik⟩+]}').:.:o	1	1	1	1	1	1	1	1	1	1	1
610	koliko	.ik.o	k}).o.:l]':ik⟩+o	1		1	1	1			1			1
611	linguistiko	.ik.o	li+nguistik⟩]}'),:.:o	1	1	1	1		1	1	1	1	1	1
612	termiko	.ik.o	te+rmi,;.:k⟩]}')o	1	1	1	1		1					1
613	entenderlo	.er.l.o	e+ntender.;.:l⟩]}')o	1	1	1	1		1	1	1	1	1	1
614	kolombiano	.an.o	kolombi⟩]}'an+,:.:o	1		1	1	1	1	1	1	1	1	1
615	mediano	.an.o	m+edi⟩a,;n]}').:o	1	1	1	1		1			1	1	1
616	peruano	.an.o	peru⟩+an]}'),;.:o	1	1	1	1	1	1	1	1	1	1	1
617	benseno	.en.o	be+ns⟩e].n}'),;:o		1	1	1		1	1	1		1	1
618	panameño	.eñ.o	panam⟩+eñ]}'),;.:o	1	1	1	1	1	1	1	1	1	1	1
619	kampesino	.in.o	ka+mpes⟩in]}'),;.:o	1	1	1	1		1	1	1	1	1	1
620	algodonero	.er.o	a+lgodon⟩er]}').:.:o	1	1	1	1		1	1	1	1	1	1
621	gayinero	.er.o	gayin⟩]}'):.er+,:o	1	1	1	1	1	1	1	1	1	1	1
622	jardinero	.er.o	ja+rdin⟩er]}'),;.:o	1	1	1	1		1	1	1	1	1	1
623	mensajero	.er.o	mensaj⟩+e:.r]}'),:o	1	1	1	1	1	1	1			1	1
624	bariyero	.iy.er.o	bariy⟩]}'),:.:er+o	1	1	1	1	1	1	1	1	1	1	1
625	kobertiso	.is.o	kober}')t⟩]is+,:.:o	1		1		1		1	1	1	1	1
626	igualito	.it.o	igu+al⟩;.it]}'),:o	1	1	1	1		1	1	1	1	1	1
627	kabayito	.it.o	ka+bay⟩]}'i)t,;.:o	1		1	1		1	1	1	1	1	1
628	kuartito	.it.o	kuart⟩]}').:.:it+o	1	1	1	1	1	1	1	1	1	1	1
629	sakatito	.it.o	saka,;.:t⟩]}')it+o	1	1	1	1	1	1					1
630	armamento	.ment.o	ar+mament⟩]}'),;.:o	1	1	1	1		1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:	
631	kolmiyo	.iy.o	ko+lm⟩]iy}'),;.:o	1	1	1	1		1	1	1	1	1	1
632	anidrido	.ido	anidr⟩]}')id,;.:o	1	1	1	1		1					1
633	amenasando	.a.ndo	a+menas⟩]}a,;.n')d:o	1		1			1	1	1	1		1
634	apartando	.a.ndo	a+part⟩]},;.:an')do	1		1			1	1	1	1	1	1
635	demandando	.a.ndo	demand⟩]}a.;.:n')d+o	1		1			1	1	1	1	1	1
636	entrando	.a.ndo	entr⟩]},;.:an')d+o	1		1			1	1	1	1	1	1
637	examinando	.a.ndo	examin⟩]}a;.n')d+,:o	1		1			1		1	1		1
638	fregando	.a.ndo	f:.reg⟩]}')and+,:o	1	1	1	1		1					1
639	kontemplando	.a.ndo	kontempl⟩]}a.n')d+,;:o	1		1			1			1		1
640	kontestando	.a.ndo	kontest⟩]}.;.:an')d+o	1		1			1	1	1	1	1	1
641	limitando	.a.ndo	limit⟩]},;.:an')d+o	1		1			1	1	1	1	1	1
642	logrando	.a.ndo	logr⟩]}a,;.:n')d+o	1		1			1	1	1	1	1	1
643	pegando	.a.ndo	p,;.:eg⟩]}an')d+o	1		1			1					1
644	preparando	.a.ndo	prepar⟩]}a,;.:n')d+o	1		1			1	1	1	1	1	1
645	yegando	.a.ndo	yeg⟩]}a,;.:n')d+o	1		1			1	1	1	1	1	1
646	absorbiendo	.ie.ndo	a+bsorb]}')iend⟩,;.:o	1	1		1		1					1
647	adkiriendo	.ie.ndo	a+dkir}i,;.:e]')nd⟩o	1	1		1		1					1
648	impidiendo	.ie.ndo	impid}ie]')nd⟩+,;.:o	1	1		1		1					1
649	kombatiendo	.ie.ndo	kombat]}.ie')n,;:d⟩+o	1	1		1		1			1		1
650	midiendo	.ie.ndo	m,;.:id}ie]')nd⟩+o	1	1		1		1					1
651	ofresiendo	.ie.ndo	o+fres]}.ie')nd⟩,;:o	1	1		1		1			1		1
652	proponiendo	.ie.ndo	pro:.pon]}')iend⟩+,:o	1	1		1		1					1
653	subiendo	.ie.ndo	su,;.:bie]}')nd⟩+o	1	1		1		1					1
654	suponiendo	.ie.ndo	su:.pon]}')iend⟩+,:o	1	1		1		1					1
655	tendiendo	.ie.ndo	t,.:e:ndie]}')nd⟩+o	1	1		1		1					1
656	uniendo	.ie.ndo	un.:i,:e]}')nd⟩+o	1	1		1		1			1	1	1
657	uyendo	.ie.ndo	u,;.:ye]n}')d⟩+o	1						1	1	1	1	
658	trineo	.eo	t,;::r.in⟩]}')e+o	1	1	1	1		1					1
659	kasikasgo	.asgo	ka+sikas]}')g⟩,;.:o											
660	despasio	.io	d],;.:es+pasi⟩}')o											
661	klasisismo	.ismo	k.lasisis]}')m⟩+,;:o											
662	nasionalismo	.ismo	nasional]}').ism⟩+,;:o	1	1		1		1			1		1
663	oportunismo	.ismo	o+portun]}'),;.:ism⟩o	1	1		1		1	1	1	1	1	1
664	positibismo	.ismo	positib]}'),;.:ism⟩+o	1	1		1		1	1	1	1	1	1
665	santanaso	.aso	sant,;.:a]}')n⟩+aso			1		1						1
666	kloridrato	.ato	kloridr⟩]}'),;.:at+o	1	1	1	1		1	1	1	1	1	1
667	sulfato	.ato	s+ulf⟩at]}'),;.:o			1								1
668	aparatito	.ito	a+parat⟩]}'),;.:ito	1	1	1	1		1	1	1	1	1	1
669	akaparamiento	.a.miento	a+kapar⟩a]}'),;.:miento	1	1	1	1		1	1	1	1	1	1
670	akatamiento	.a.miento	a+k,;.:at⟩a]}')miento	1	1	1	1		1					1
671	akondisionamiento	.a.miento	a,;.:kondision⟩}')a]miento	1	1	1	1		1					1
672	akoplamiento	.a.miento	a+kopl⟩}')a],;.:miento	1	1	1	1		1	1	1	1	1	1
673	enkarselamiento	.a.miento	enkarsel⟩+a]}'),;.:miento	1	1	1	1	1	1	1	1	1	1	1
674	planeamiento	.a.miento	plane⟩}').a]mient+,::o	1	1	1	1		1			1		1
675	atrebimiento	.i.miento	a+trebi]}')mient⟩,;.:o	1	1		1		1					1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	.	:	.	:	
676	deskubrimiento	.i.miento	deskubr}'i]m)ient⟩+.:.:o	1			1		1					
677	dibertimiento	.i.miento	dibert)i]}',:.:mient⟩+o	1	1		1		1	1	1	1	1	1
678	konbensimiento	.i.miento	konben,:.:si]}')mient⟩+o	1	1		1		1					1
679	padesimiento	.i.miento	pades]}i')mient⟩+.:.:o	1	1		1		1					1
680	surjimiento	.i.miento	s,:.:urji]}'m)ient⟩+o	1			1		1					
681	kampamento	.amento	ka+mp]}')ament⟩,:.:o	1	1		1		1					1
682	abrasar	.a.r	a+bras⟩]}')a.:.:r	1	1	1	1		1	1	1	1	1	1
683	actualisar	.a.r	a+ctualis⟩a]}'),:.:r	1	1	1	1		1	1	1	1	1	1
684	afrontar	.a.r	afron+t⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
685	aguantar	.a.r	a+guant⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
686	akreditar	.a.r	a+kredit⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
687	apelar	.a.r	a+pel⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
688	apuntar	.a.r	apun+t⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
689	atakar	.a.r	a+tak⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
690	autorisar	.a.r	a+utoris⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
691	ayudar	.a.r	a+yud⟩]}')a.:.:r	1	1	1	1		1	1	1	1	1	1
692	berifikar	.a.r	berifik⟩+]}')a.:.:r	1	1	1	1	1	1	1	1	1	1	1
693	brindar	.a.r	b+rind⟩]}')a.:.:r	1	1	1	1		1	1	1	1	1	1
694	deklarar	.a.r	deklar⟩+]}')a,:.:r	1	1	1	1	1	1	1	1	1	1	1
695	deklinar	.a.r	deklin⟩+]}')a,:.:r	1	1	1	1	1	1	1	1	1	1	1
696	detectar	.a.r	detec+t⟩]}')a.:.:r	1	1	1	1		1	1	1	1	1	1
697	ensamblar	.a.r	en+sambl⟩}')a],:.:r	1	1	1	1		1	1	1	1	1	1
698	enunsiar	.a.r	e.:.:n+unsi⟩}')a]r	1	1	1	1		1					1
699	figurar	.a.r	fig+ur⟩]}'),:a.:r	1	1	1	1		1	1	1	1	1	1
700	implikar	.a.r	implik⟩+]}'):a,.:r	1	1	1	1	1	1	1	1	1	1	1
701	improbisar	.a.r	imp+robis⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
702	impulsar	.a.r	imp+uls⟩]}'):.a,:r	1	1	1	1		1	1	1	1	1	1
703	juntar	.a.r	jun+t⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
704	kolaborar	.a.r	kolab+or⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
705	komunikar	.a.r	komunik⟩+]}')a,:.:r	1	1	1	1	1	1	1	1	1	1	1
706	kontestar	.a.r	kontest⟩+]}')a,:.:r	1	1	1	1	1	1	1	1	1	1	1
707	kosinar	.a.r	kosin⟩+]}a'),:.:r	1	1	1	1	1	1	1	1	1	1	1
708	kritikar	.a.r	kritik⟩+]}')a,:.:r	1	1	1	1	1	1	1	1	1	1	1
709	lebantar	.a.r	le+bant⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
710	legalisar	.a.r	le+galis⟩a]}'),:.:r	1	1	1	1		1	1	1	1	1	1
711	molekular	.a.r	molekul⟩+a]}'),:.:r	1	1	1	1	1	1	1	1	1	1	1
712	nombrar	.a.r	nomb+r⟩]}'),:.:ar	1	1	1	1		1	1	1	1	1	1
713	opinar	.a.r	o+pin⟩]}'),:.:ar	1	1	1	1		1	1	1	1	1	1
714	presensiar	.a.r	pre+sensi⟩]}')a,:.:r	1	1	1	1		1	1	1	1	1	1
715	sustentar	.a.r	sustent⟩+]}')a,:.:r	1	1	1	1	1	1	1	1	1	1	1
716	tolerar	.a.r	toler⟩+]}')a,:.:r	1	1	1	1	1	1	1	1	1	1	1
717	transpeninsular	.a.r	trans,:.:peninsul⟩+}ar			1		1	1					1
718	umiyar	.a.r	umiy⟩+]}')a,:.:r	1	1	1	1	1	1	1	1	1	1	1
719	akonteser	.e.r	a+kontes⟩,:.:e]}')r	1	1	1	1		1	1	1	1	1	1
720	apreender	.e.r	a+pre.end⟩e]}'),:.:r	1	1	1	1		1	1	1		1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	:	.	:	
721	eskojer	.e.r	e+s,;.:koj⟩e]}')r	1	1	1	1		1					1
722	fortaleser	.e.r	fortal+es⟩.e]}'),;:r	1	1	1	1		1	1	1	1	1	1
723	kareser	.e.r	k..:a:res⟩+e]}')r	1	1	1	1	1	1					1
724	kreser	.e.r	k,.:re:s⟩+e]}')r	1	1	1	1	1	1					1
725	perteneser	.e.r	per+tenes⟩e]}').:.:r	1	1	1	1		1	1	1	1	1	1
726	aplaudir	.i.r	a+plaud⟩i]}'),;.:r	1	1	1	1		1	1	1	1	1	1
727	inkluir	.i.r	inklu⟩+i]}'),;.:r	1	1	1	1	1	1	1	1	1	1	1
728	mentir	.i.r	me+nt⟩i]}'),;.:r	1	1	1	1		1	1	1	1	1	1
729	nutrir	.i.r	nu+tr⟩)i]}',;.:r	1	1	1	1		1	1	1	1	1	1
730	presidir	.i.r	p+resid⟩,;.i]}'):r	1	1	1	1		1	1	1	1	1	1
731	suplir	.i.r	su+pl⟩)i]}',;.:r	1	1	1	1		1	1	1	1	1	1
732	surjir	.i.r	s+urj⟩i]}'),;.:r	1	1	1	1		1	1	1	1	1	1
733	abentar	.ar	a+b,;.:ent⟩]}')ar	1	1	1	1		1					1
734	abitar	.ar	a+.:bi,;t⟩]}')ar	1	1	1	1		1					1
735	apagar	.ar	a,;.:pag⟩+]}')ar	1	1	1	1	1	1					1
736	ensayar	.ar	e+nsay⟩]}'),;.:ar	1	1	1	1		1	1	1	1	1	1
737	entonar	.ar	en+to,;n⟩]}').:ar	1	1	1	1		1			1	1	1
738	frenar	.ar	fren⟩+]}'),;.:ar	1	1	1	1	1	1	1	1	1	1	1
739	klabar	.ar	k..:lab⟩+]}');ar	1	1	1	1	1	1		1			1
740	kostar	.ar	ko').:st⟩+]},;ar	1		1		1	1	1	1			1
741	lesionar	.ar	le,;.:sion⟩+]}')ar	1	1	1	1	1	1					1
742	presionar	.ar	pre,;.:sion⟩+]}')ar	1	1	1	1	1	1					1
743	sentrar	.ar	s.:e+nt,;r⟩]}')ar	1	1	1	1		1					1
744	soltar	.ar	s+lo,;.:t⟩]}')ar	1	1	1	1		1					1
745	unifikar	.ifik.ar	un,;.:i+fik⟩]}')ar	1	1	1	1		1	1	1	1	1	1
746	inskribir	.ir	in,;.:skrib⟩+]}')ir	1	1	1	1	1	1					1
747	bector	.or	b]}.:e,:ct+or					1						
748	eskultor	.or	e+skul⟩)t]}',;.:or	1			1		1	1	1	1	1	
749	posterior	.or	p+os⟩}teri]'o,;.:r	1			1							
750	sensor	.or	se+n⟩s.:o]}'),;r									1	1	
751	aparador	.a.dor	a+,;.:par⟩]}')ador	1	1	1	1		1					1
752	aplikador	.a.dor	a+plik⟩ado]}'),;.:r			1								1
753	aselerador	.a.dor	a+seler⟩ado]}'),;.:r			1								1
754	asosiador	.a.dor	a+sosi⟩ado]}'),;.:r			1								1
755	donador	.a.dor	don⟩+,;.:ado]}')r			1		1		1	1	1	1	1
756	inbestigador	.a.dor	in+bestig⟩ado]}'),;.:r			1								1
757	indikador	.a.dor	indik⟩+ado]}'),;.:r			1		1						1
758	istoriador	.a.dor	is+tori⟩ado]}'),;.:r			1								1
759	matador	.a.dor	ma,;.:t⟩a+do]}')r			1		1						1
760	entras	.a.s	ent,;r⟩+}').:a]s	1	1	1	1	1	1			1	1	1
761	lakras	.a.s	l,;:a.k]}')r⟩+as			1		1						1
762	likidas	.a.s	li+kid⟩}').:a],;s	1	1	1	1		1	1	1	1	1	1
763	mandas	.a.s	m,;and⟩+}')a].:s	1	1	1	1	1	1			1	1	1
764	muestras	.a.s	muest+r⟩a]}'),;.:s	1	1	1	1		1	1	1	1	1	1
765	organikas	.a.s	o+rganik⟩}a]'),;.:s	1	1	1	1		1	1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	:	.	:	
766	trankilas	.a.s	tra+nkil)}.:a]').:s	1	1	1	1		1	1	1	1	1	1
767	moldeadas	.ad.a.s	mo+lde⟩;.ad]}'),:as	1	1	1	1		1	1	1	1	1	1
768	abajeñas	.eñ.a.s	a+baj]}'),;.:eñ⟩as	1	1	1	1		1	1	1	1	1	1
769	ponedoras	.e.dor.a.s	pone].dor⟩+}'),;:as	1		1	1	1	1	1	1	1	1	1
770	konsiensias	.ia.s	konsiensi⟩+.:a]}').:s	1	1		1		1			1	1	1
771	progresistas	.ista.s	pro+gresist⟩a]}'),;.:s	1	1		1		1	1	1	1	1	1
772	perdones	.e.s	p+erd.:on⟩:e]}').s	1	1	1	1		1			1	1	1
773	series	.e.s	se+r⟩i.:e]}').:s	1	1		1		1	1	1	1	1	1
774	sobres	.e.s	sob.:r⟩+e]}').:s	1	1	1	1	1	1			1	1	1
775	personajes	.aje.s	per+son⟩aje]}'),;.:s	1	1	1	1		1	1	1	1	1	1
776	belijerantes	.a.nte.s	belijer)]}'),;.:ant+es	1	1	1	1		1	1	1	1	1	1
777	exitos	.o.s	e+x⟩it.:o]}').:s	1	1		1		1	1	1	1	1	1
778	kargos	.o.s	ka+rg)]},;.:o')s	1	1	1	1		1	1	1	1	1	1
779	pingos	.o.s	p.;.:i+ng)]}')os	1	1	1	1		1					1
780	elados	.ad.o.s	el⟩.:ad+),;o]}'s	1	1	1	1	1	1	1	1	1	1	1
781	pekados	.ad.o.s	pek⟩ad+o]}'),;.:s	1	1	1	1	1	1	1	1	1	1	1
782	bebedisos	.is.o.s	bebed⟩]}'.;.:is+os	1		1	1	1	1	1	1	1	1	1
783	tamalitos	.it.o.s	ta.;.:mal⟩]}')it+os	1	1	1	1	1	1					1
784	feroses	.es	f+er⟩os]}'),;.:es	1	1		1		1	1	1	1	1	1
785	fertiles	.es	fe+rt⟩il]}'),;.:es	1	1		1		1	1	1	1	1	1
786	jabones	.es	jab]}').on⟩,;:e+s			1				1	1		1	1
787	ostiles	.es	o+st⟩il]}'),;.:es	1	1		1		1	1	1	1	1	1
788	paredes	.es	p+ar,;.:ed⟩]}')es	1	1	1	1		1					1
789	pormenores	.es	po+r.;.::menor)]}')es	1	1	1	1		1					1
790	fundamentales	.al.es	fu+ndament⟩al]}'),;.:es	1	1	1	1		1	1	1	1	1	1
791	komersiales	.al.es	ko+mersi⟩al]}'),;.:es	1	1	1	1		1	1	1	1	1	1
792	kordiales	.al.es	ko+rdi⟩al]}'),;.:es	1	1	1	1		1	1	1	1	1	1
793	liberales	.al.es	libe+r⟩al]}'),;.:es	1	1	1	1		1	1	1	1	1	1
794	matrimoniales	.al.es	ma+trimoni⟩al]}'),;.:es	1	1	1	1		1	1	1	1	1	1
795	mediebales	.al.es	medieb⟩+al]}'),;.:es	1	1	1	1	1	1	1	1	1	1	1
796	ofisiales	.al.es	ofis+i⟩al]}'),;.:es	1	1	1	1		1	1	1	1	1	1
797	teatrales	.al.es	te+atr⟩al]}'),;.:es	1	1	1	1		1	1	1	1	1	1
798	jentiles	.il.es	j+ent⟩i')l]};.e,:s	1		1			1		1	1		1
799	tayarines	.in.es	ta+yar⟩]}')in.;.:es	1	1	1	1		1	1	1	1	1	1
800	peones	.on.es	p;e,.:on⟩]}')e+s	1	1	1	1		1	1		1	1	1
801	instrucsiones	.sion.es	i,;.:nstruc]}si')on⟩+es	1		1		1	1					1
802	lecsiones	.sion.es	l,;.:ecsi]}')on⟩+es			1		1						1
803	lokusiones	.sion.es	lo,;.:ku]si}')on⟩+es	1		1		1						1
804	perkusiones	.sion.es	p,;.:e+rkusi]}on⟩')es		1	1	1							1
805	satisfacsiones	.sion.es	sa+tisfac,;.:si]}')on⟩es			1				1	1	1	1	1
806	transisiones	.sion.es	trans.i]},;:si')on⟩+es	1		1		1	1	1	1		1	1
807	ekibokasiones	.a.sion.es	e+kibok)]}a,;.:si')ones	1		1			1	1	1	1	1	1
808	komparasiones	.a.sion.es	kompar⟩]}a,;.:si')on+es	1		1		1	1	1	1	1	1	1
809	konstelasiones	.a.sion.es	konst.;.:el⟩+asi]}')ones			1		1						1
810	korporasiones	.a.sion.es	kor+por⟩a,;.:si]}')ones			1				1	1	1	1	1

Tabla B.4 (continuación):
Muestra aletoria de vocablos analizados

	vocablo	sufs.	segmentaciones	]	)	⟩	'	+	}	,	;	.	:	
811	lokasiones	.a.sion.es	l.:ok⟩]}a,;si')on+es	1		1		1	1	1	1			1
812	notifikasiones	.a.sion.es	no,:.:tifik⟩]}asi')on+es	1		1		1	1					1
813	partisipasiones	.a.sion.es	pa+rtisip⟩]}a,;.:si')ones	1		1			1	1	1	1	1	1
814	perforasiones	.a.sion.es	per+for⟩]}a,;.:si')ones	1		1			1	1	1	1	1	1
815	preokupasiones	.a.sion.es	preokup⟩]}a,;.:si')on+es	1		1		1	1	1	1	1	1	1
816	publikasiones	.a.sion.es	pu+blik⟩]}.:.:asi')ones	1		1			1	1	1	1	1	1
817	berases	.as.es	b+er⟩,;as]}')).:es	1	1	1	1		1	1	1	1	1	1
818	institutrises	.tris.es	in+stitu,;.:tr⟩]}')ises							1	1	1	1	
819	atabales	.aba.les	a+tab⟩,;.:a]}')les	1	1		1		1					1
820	diagnosis	.is	dia+gnos⟩]}',;.:is	1	1	1	1		1	1	1	1	1	1
821	beamos	.a.mos	be⟩]},;.:a')m+os	1	1	1	1		1	1	1	1	1	1
822	jeneralisamos	.a.mos	jeneralis⟩')a]},;.:m+os	1	1	1	1		1	1	1	1	1	1
823	kondenamos	.a.mos	konden⟩}')a],;.:m+os	1	1	1	1		1	1	1	1	1	1
824	pidamos	.a.mos	pid⟩]},;.:a')m+os	1	1	1	1		1	1	1	1	1	1
825	pongamos	.a.mos	p,;.:ong⟩}')a]m+os	1	1	1	1		1					1
826	senamos	.a.mos	se,n⟩]}').:a;m+os	1	1	1	1		1		1	1	1	1
827	ibamos	.ba.mos	i,;.:b⟩a]}')m+os	1	1		1		1	1	1	1	1	1
828	andabamos	.aba.mos	a+nd⟩aba]}'),;.:mos	1	1	1	1		1	1	1	1	1	1
829	dieramos	.ie.ra.mos	d.:i,:er⟩})a]'m+os	1			1					1	1	
830	atrebemos	.e.mos	a+trebe]}'),;.:m⟩os	1	1		1		1	1	1	1	1	1
831	yegemos	.e.mos	yeg]},;.:e')m⟩+os	1	1		1		1	1	1	1	1	1
832	abrimos	.i.mos	abir]}'),;.:m⟩+os	1	1		1		1	1	1	1	1	1
833	asistimos	.i.mos	a,;:sisti]}').m⟩+os	1	1		1		1			1		1
834	desidimos	.i.mos	desidi]}'),;.:m⟩+os	1	1		1		1	1	1	1	1	1
835	eskribimos	.i.mos	eskribi]}'),;.:m⟩+os	1	1		1		1	1	1	1	1	1
836	yegaremos	.a.r.emos	yeg⟩}ar],;.:e')m+os	1		1			1	1	1	1	1	1
837	benseremos	.e.r.emos	b,;:ense]}')r.em⟩+os	1	1		1		1			1		1
838	deberemos	.er.emos	de,.:ber]};e')m⟩+os	1					1		1			
839	dandonos	.a.ndo.nos	d),;.:ando]}'n⟩+os	1	1		1		1	1	1	1	1	1
840	ayudarnos	.ar.nos	a+yud⟩]}')ar,;.:nos	1	1	1	1		1	1	1	1	1	1
841	separarnos	.ar.nos	se+par⟩]}')ar,;.:nos	1	1	1	1		1	1	1	1	1	1
842	beintisiete	.i.siete	beint]'i,;.:si⟩+}ete	1			1			1	1	1	1	
843	beintiuna	.i.un.a	be+inti;.un⟩]}'),:a	1	1	1	1		1	1	1	1	1	1
844	diesisiete	.i.siete	di+esi,;.:si⟩]}')ete							1	1	1	1	
845	koliflor	.i.flor	kol+i,;.:f⟩}lor					1		1	1	1	1	

Apéndice C

Sufijos en el *CEMC*

En este apéndice se listan los 749 fragmentos de vocablos gráficos más afijales (según las medidas para segmentarlos aplicadas en el capítulo dedicado al afijo) y algunos sufijos derivativos identificables por su forma (que no se examinaron en ese capítulo).

Los 749 fragmentos se presentan en la tabla C.1 en orden del más al menos afijal. La medida de afijalidad (columna 9) es el promedio de los índices normalizados de cantidades de cuadros, economía y entropía (columnas 4-6) calculados para cada segmento a partir del *CEMC*, según los criterios expuestos en dicho capítulo. La frecuencia de los sufijos (columna 3) no se refiere a la mera ocurrencia del fragmento en cualquier vocablo, sino solamente a su ocurrencia como afijo, según los mismos criterios. Nótese, además, que dicha frecuencia no necesariamente implica mayor afijalidad, es decir, si bien los segmentos de vocablo más frecuentes tienden a ser los más afijales, ni el más frecuente es el más afijal, ni los menos frecuentes se excluyen ni ocurren necesariamente al final (a menos que su afijalidad sea baja). Por último, las probabilidades 1 y 2 (columnas 7 y 8), cuyos valores mayores tampoco implican necesariamente mayor afijalidad, se refieren a las ocurrencias de los afijos ya sea en

el corpus, Ψ, o en la lista de vocablos, Φ (véase formalización a partir de la página 124 y discusión a partir de la 135).

Al final de este apéndice, en la tabla C.2 (a partir de la página 419) se listan algunos sufijos de derivación nominal que se identificaron en la tabla C.1, pero que por falta de espacio no se discutieron en el capítulo sobre el afijo. Los sufijos están agrupados improvisadamente por significado y, sobre todo, por semejanza formal.

Tabla C.1: Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
1.	~ó	1428	0.7371	0.9192	0.872	0.8745	0.9003	0.8428
2.	~o	6314	0.686	0.9788	0.8017	0.4695	0.6291	0.8222
3.	~s	12013	1	0.9968	0.4609	0.5378	0.5125	0.8192
4.	~a	7687	0.5753	0.9818	0.8888	0.5153	0.4431	0.8153
5.	~os	4554	0.4775	0.9754	0.8235	0.5162	0.5639	0.7588
6.	~as	4324	0.4216	0.9779	0.8645	0.6075	0.5965	0.7547
7.	~en	945	0.4107	0.8991	0.906	0.863	0.2368	0.7386
8.	~ar	1633	0.2178	0.9621	0.9149	0.7346	0.8928	0.6982
9.	~ado	1429	0.2061	0.9619	0.907	0.7099	0.9231	0.6917
10.	~ando	976	0.1836	0.9544	0.9162	0.8399	0.9708	0.6847
11.	~e	2363	0.42	0.9482	0.6817	0.2738	0.2295	0.6833
12.	~é	639	0.4104	0.8198	0.8153	0.8925	0.409	0.6818
13.	~aba	828	0.1821	0.9565	0.9024	0.8894	0.9564	0.6803
14.	~aron	736	0.1779	0.9604	0.8935	0.8943	0.9726	0.6773
15.	~ada	1135	0.1654	0.9491	0.9159	0.7385	0.9227	0.6768
16.	~arse	665	0.1462	0.9541	0.9072	0.8428	0.9521	0.6692
17.	~ados	941	0.1477	0.9549	0.9008	0.7189	0.8582	0.6678
18.	~aban	551	0.1434	0.9395	0.9002	0.9062	0.9578	0.661
19.	~adas	813	0.1316	0.9449	0.9041	0.767	0.8687	0.6602
20.	~an	1775	0.195	0.9434	0.8354	0.6187	0.6729	0.6579
21.	~ara	370	0.1098	0.9151	0.9151	0.8916	0.9848	0.6467
22.	~ará	387	0.121	0.9295	0.8739	0.9214	0.9021	0.6415
23.	~arlo	316	0.09269	0.9291	0.8849	0.9159	0.9588	0.6356
24.	~arla	270	0.0795	0.9185	0.9071	0.931	0.965	0.635
25.	~arme	244	0.08683	0.9134	0.8916	0.9313	0.9537	0.6306
26.	~andose	260	0.07949	0.9136	0.8855	0.8966	0.9067	0.6262
27.	~arán	256	0.08995	0.9112	0.8759	0.9242	0.9118	0.6257
28.	~ido	445	0.1038	0.8567	0.914	0.7672	0.8516	0.6248
29.	~ita	453	0.09631	0.8965	0.8729	0.7639	0.8077	0.6219
30.	~aría	231	0.08063	0.8869	0.892	0.924	0.9059	0.6198
31.	~amos	645	0.1345	0.8801	0.8415	0.738	0.8804	0.6187
32.	~amente	624	0.1189	0.9784	0.7534	0.8607	0.9793	0.6169
33.	~arlos	201	0.06461	0.8959	0.8853	0.9095	0.9235	0.6153
34.	~ador	268	0.05489	0.8927	0.8965	0.7768	0.7696	0.6147
35.	~aran	196	0.07454	0.8688	0.8857	0.9245	0.9145	0.6097
36.	~es	2479	0.1876	0.9529	0.6885	0.4846	0.6193	0.6097
37.	~ito	421	0.09323	0.8752	0.8594	0.736	0.7269	0.6093
38.	~aste	136	0.05731	0.8344	0.9237	0.8662	0.8713	0.6051
39.	~arte	144	0.05526	0.8446	0.9136	0.9057	0.8802	0.6045
40.	~antes	187	0.03646	0.8802	0.8944	0.6404	0.4992	0.6037

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
41.	~adores	196	0.04172	0.8653	0.9029	0.7717	0.8551	0.6033
42.	~idos	269	0.07274	0.8321	0.9042	0.727	0.8084	0.603
43.	~ida	304	0.08306	0.8372	0.8864	0.7221	0.9108	0.6022
44.	~arlas	139	0.05346	0.8775	0.8755	0.891	0.9136	0.6021
45.	~asión	540	0.06338	0.9278	0.8111	0.5273	0.8398	0.6007
46.	~amiento	141	0.02085	0.8871	0.8935	0.6104	0.8256	0.6005
47.	~ante	240	0.03402	0.8769	0.8883	0.6383	0.7293	0.5998
48.	~arle	176	0.06573	0.8618	0.8716	0.9514	0.9866	0.5997
49.	~asiones	263	0.04548	0.9155	0.8365	0.6726	0.8612	0.5992
50.	~aremos	104	0.04195	0.8431	0.9103	0.9369	0.8746	0.5985
51.	~itos	271	0.05928	0.8659	0.8699	0.7188	0.6811	0.5984
52.	~n	3586	0.2767	0.9634	0.553	0.445	0.213	0.5977
53.	~asas	9	0.01902	0.935	0.8378	0.3	0.3375	0.5973
54.	~itas	210	0.05322	0.8721	0.8633	0.7071	0.6158	0.5962
55.	~aré	127	0.06152	0.8274	0.8992	0.8944	0.8278	0.596
56.	~ero	292	0.04169	0.846	0.8981	0.6838	0.8634	0.5953
57.	~ir	209	0.05132	0.8624	0.8675	0.6875	0.7086	0.5938
58.	~aja	10	0.0161	0.9146	0.8505	0.3125	0.4606	0.5937
59.	~ase	114	0.0352	0.8362	0.9062	0.6706	0.4036	0.5925
60.	~arnos	139	0.05347	0.8634	0.8582	0.8854	0.9387	0.5917
61.	~oro	15	0.02238	0.9094	0.8383	0.3333	0.4809	0.59
62.	~eros	200	0.03397	0.8269	0.8991	0.627	0.7706	0.5867
63.	~años	7	0.01745	0.9471	0.7924	0.35	0.03583	0.5857
64.	~eso	17	0.02764	0.8713	0.8496	0.2698	0.2388	0.5828
65.	~eras	137	0.02477	0.7897	0.9322	0.548	0.58	0.5822
66.	~alas	14	0.01541	0.8617	0.8681	0.4	0.6756	0.5817
67.	~adora	124	0.03011	0.8301	0.8829	0.7086	0.7482	0.581
68.	~idas	218	0.06886	0.8143	0.8598	0.7101	0.7775	0.581
69.	~osa	178	0.02319	0.8343	0.8831	0.522	0.7175	0.5802
70.	~usa	10	0.01146	0.8607	0.8631	0.3333	0.1708	0.5784
71.	~ilo	14	0.01852	0.8589	0.8518	0.2692	0.2815	0.5764
72.	~abamos	115	0.05097	0.8073	0.8657	0.92	0.9128	0.5746
73.	~iendo	276	0.103	0.8603	0.758	0.899	0.9201	0.5738
74.	~anta	16	0.01239	0.8446	0.863	0.5517	0.7335	0.5733
75.	~oso	191	0.02696	0.8122	0.8804	0.5685	0.5802	0.5732
76.	~ala	30	0.01649	0.8084	0.8931	0.4918	0.7365	0.5727
77.	~orar	6	0.005559	0.8422	0.8682	0.2	0.3463	0.572
78.	~adoras	41	0.01006	0.8282	0.8765	0.5694	0.5548	0.5716
79.	~arían	81	0.0356	0.8054	0.8734	0.9205	0.9107	0.5715
80.	~amo	7	0.01011	0.9026	0.798	0.2593	0.2917	0.5703

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
81.	~ió	303	0.09638	0.8693	0.7438	0.7769	0.888	0.5698
82.	~ame	74	0.03199	0.7514	0.9228	0.7475	0.7989	0.5687
83.	~ana	58	0.006222	0.8263	0.8732	0.2613	0.2582	0.5686
84.	~ija	7	0.009551	0.8593	0.8359	0.2917	0.07743	0.5682
85.	~ayo	9	0.01594	0.8619	0.8264	0.3103	0.4622	0.5681
86.	~i	115	0.01817	0.6859	1	0.2854	0.03225	0.568
87.	~able	115	0.02224	0.8142	0.8673	0.4423	0.5819	0.5679
88.	~osos	112	0.01852	0.8169	0.8664	0.5185	0.4777	0.5673
89.	~esos	8	0.0266	0.9658	0.7093	0.2286	0.4486	0.5672
90.	~ijo	10	0.01716	0.8747	0.8088	0.3704	0.7541	0.5669
91.	~alado	6	0.02259	0.8586	0.8194	0.2857	0.07767	0.5669
92.	~ases	19	0.008563	0.8774	0.813	0.475	0.6752	0.5664
93.	~er	264	0.07202	0.7753	0.8509	0.6241	0.5463	0.5661
94.	~apa	9	0.01506	0.9193	0.7636	0.4737	0.5	0.566
95.	~irse	108	0.03451	0.8176	0.8456	0.777	0.7351	0.5659
96.	~ata	32	0.01854	0.7699	0.9076	0.2991	0.7195	0.5653
97.	~osas	91	0.01128	0.825	0.8597	0.474	0.7957	0.5653
98.	~istas	181	0.0225	0.8434	0.8295	0.6704	0.7315	0.5651
99.	~andolo	78	0.0264	0.815	0.8507	0.8478	0.8644	0.564
100.	~alo	45	0.01486	0.8002	0.8766	0.5625	0.8047	0.5639
101.	~ate	107	0.03597	0.7587	0.8966	0.7133	0.7581	0.5637
102.	~ismo	213	0.02256	0.8756	0.7915	0.6034	0.8445	0.5632
103.	~eto	10	0.003645	0.8382	0.8459	0.1613	0.08539	0.5626
104.	~osar	8	0.007829	0.92	0.7584	0.6154	0.8214	0.5621
105.	~olar	6	0.001196	0.8727	0.811	0.24	0.5466	0.5616
106.	~eses	26	0.01229	0.8106	0.8609	0.4561	0.9426	0.5613
107.	~ilado	6	0.027	0.8068	0.8498	0.2727	0.3043	0.5612
108.	~ables	88	0.02402	0.828	0.8312	0.44	0.4916	0.5611
109.	~asta	9	0.01317	0.8704	0.7982	0.45	0.05859	0.5606
110.	~ika	288	0.01552	0.9096	0.7503	0.375	0.3802	0.5585
111.	~andola	58	0.02131	0.761	0.8929	0.8169	0.828	0.5584
112.	~iko	341	0.01831	0.893	0.7629	0.4231	0.5631	0.5581
113.	~anes	29	0.002902	0.8811	0.7901	0.5918	0.8776	0.558
114.	~una	10	0.02008	0.7951	0.8589	0.303	0.01535	0.558
115.	~asa	21	0.01497	0.8285	0.8306	0.28	0.8781	0.558
116.	~aya	13	0.01304	0.7669	0.894	0.3171	0.49	0.558
117.	~anas	18	0.003933	0.85	0.8193	0.1579	0.2032	0.5577
118.	~era	335	0.03686	0.7969	0.8368	0.6025	0.5334	0.5569
119.	~ista	231	0.02466	0.8414	0.8005	0.6834	0.8885	0.5555
120.	~arás	46	0.02251	0.7748	0.8689	0.7797	0.6848	0.5554

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
121.	~iso	28	0.02007	0.7573	0.8886	0.4	0.2261	0.5553
122.	~eje	10	0.02068	0.8118	0.8333	0.4762	0.495	0.5553
123.	~esas	24	0.01534	0.7855	0.8645	0.3333	0.08998	0.5551
124.	~ikas	167	0.009263	0.9093	0.7467	0.3902	0.3599	0.5551
125.	~orado	6	0.004105	0.8256	0.835	0.15	0.1717	0.5549
126.	~abas	30	0.01624	0.784	0.8629	0.5172	0.5475	0.5544
127.	~ikos	193	0.0117	0.917	0.7345	0.3621	0.4359	0.5544
128.	~ieron	238	0.08253	0.8352	0.7448	0.9189	0.9736	0.5542
129.	~oja	7	0.005147	0.8844	0.7726	0.28	0.183	0.554
130.	~ansas	14	0.003519	0.783	0.871	0.4516	0.36	0.5525
131.	~ee	9	0.01569	0.7832	0.8557	0.2143	0.5394	0.5515
132.	~al	375	0.02833	0.8187	0.8073	0.4584	0.3521	0.5515
133.	~andome	48	0.0209	0.7444	0.8883	0.7619	0.7975	0.5512
134.	~ales	281	0.02451	0.8433	0.7848	0.4973	0.6488	0.5509
135.	~amientos	47	0.009897	0.7919	0.8499	0.6812	0.6164	0.5506
136.	~arles	72	0.03206	0.8179	0.8015	0.9351	0.9572	0.5505
137.	~ino	48	0.005806	0.791	0.8532	0.381	0.7858	0.55
138.	~inas	38	0.004263	0.7719	0.8732	0.25	0.4483	0.5498
139.	~asen	27	0.01459	0.7357	0.8988	0.587	0.106	0.5497
140.	~aso	94	0.01204	0.7555	0.8812	0.6528	0.8431	0.5496
141.	~us	41	0.007607	0.729	0.9117	0.3178	0.014	0.5494
142.	~ano	67	0.009186	0.7886	0.85	0.2735	0.4499	0.5493
143.	~ale	48	0.02016	0.7096	0.9174	0.6575	0.8889	0.549
144.	~udo	21	0.007024	0.7254	0.9146	0.3443	0.4894	0.549
145.	~idor	12	0.004152	0.7965	0.8455	0.3871	0.2171	0.5487
146.	~ina	115	0.01349	0.7588	0.8728	0.3117	0.3959	0.5483
147.	~irte	8	0.02217	0.7742	0.8432	0.2857	0.495	0.5465
148.	~etos	6	0.01438	0.8037	0.8212	0.1364	0.1106	0.5464
149.	~ieras	7	0.003418	0.9273	0.7078	0.2258	0.6218	0.5462
150.	~alos	32	0.01279	0.7583	0.8663	0.5333	0.7545	0.5458
151.	~adamente	53	0.0101	0.8126	0.8144	0.6163	0.7825	0.5457
152.	~ona	95	0.01317	0.7955	0.8273	0.5398	0.7575	0.5453
153.	~ila	16	0.00227	0.7306	0.9027	0.3019	0.2762	0.5452
154.	~ako	10	0.002801	0.8434	0.789	0.2273	0.4448	0.5451
155.	~aras	28	0.01889	0.7032	0.9117	0.4118	0.6319	0.5446
156.	~adero	16	0.006166	0.7386	0.8868	0.4324	0.1348	0.5439
157.	~atibo	55	0.01047	0.8575	0.7575	0.6962	0.8503	0.5418
158.	~andole	75	0.03218	0.7533	0.8395	0.9036	0.8889	0.5417
159.	~año	9	0.01478	0.8112	0.7989	0.3462	0.1465	0.5416
160.	~igo	12	0.0123	0.8331	0.7785	0.2791	0.652	0.5413

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
161.	~eno	15	0.01444	0.8141	0.7953	0.1456	0.06686	0.5413
162.	~uto	9	0.003096	0.738	0.882	0.2368	0.07843	0.541
163.	~ojo	6	0.004762	0.8198	0.7978	0.25	0.1243	0.5408
164.	~ear	75	0.01772	0.7835	0.8202	0.4808	0.4505	0.5405
165.	~isos	11	0.01565	0.7648	0.8409	0.3056	0.4409	0.5405
166.	~etas	31	0.013	0.7547	0.852	0.301	0.2737	0.5399
167.	~ete	59	0.01628	0.6704	0.9328	0.4436	0.4179	0.5398
168.	~anos	53	0.007624	0.7827	0.8284	0.3011	0.4617	0.5396
169.	~rando	10	0.002632	0.8757	0.74	0.07874	0.06076	0.5394
170.	~uro	16	0.01504	0.7569	0.8449	0.3265	0.5935	0.539
171.	~urar	6	0.003026	0.8358	0.778	0.1622	0.5535	0.5389
172.	~ten	31	0.005389	0.8487	0.7627	0.1987	0.2352	0.5389
173.	~tado	54	0.006453	0.8897	0.7195	0.1824	0.3497	0.5385
174.	~lo	792	0.08679	0.9323	0.596	0.7084	0.2755	0.5384
175.	~atiba	44	0.008576	0.8427	0.7633	0.6471	0.8644	0.5382
176.	~laba	7	0.02133	0.8208	0.772	0.1429	0.342	0.538
177.	~ota	24	0.007512	0.7076	0.8974	0.3582	0.5209	0.5375
178.	~ía	970	0.1033	0.82	0.688	0.7602	0.7894	0.5371
179.	~ačo	6	0.001712	0.8736	0.7356	0.3	0.1	0.537
180.	~ato	60	0.009219	0.7489	0.8527	0.4225	0.4815	0.5369
181.	~esa	51	0.01987	0.6808	0.9099	0.3806	0.1738	0.5368
182.	~atos	20	0.008973	0.8059	0.7946	0.2532	0.6438	0.5365
183.	~ería	121	0.02834	0.7	0.881	0.6173	0.8725	0.5364
184.	~enas	10	0.004729	0.8017	0.8008	0.2174	0.06921	0.5357
185.	~és	89	0.01533	0.7229	0.8688	0.5361	0.5744	0.5357
186.	~andolas	22	0.007984	0.7641	0.8348	0.6111	0.5897	0.5356
187.	~inos	34	0.005286	0.7818	0.8197	0.3696	0.2997	0.5356
188.	~eo	99	0.01667	0.7453	0.8407	0.4249	0.6282	0.5342
189.	~ajes	20	0.003786	0.7302	0.8681	0.3448	0.2487	0.534
190.	~se	1619	0.1485	0.9411	0.51	0.7615	0.2373	0.5332
191.	~iera	165	0.05362	0.7793	0.7667	0.8967	0.9379	0.5332
192.	~istes	6	0.003472	0.8306	0.7647	0.2727	0.5361	0.5329
193.	~taron	29	0.008717	0.8474	0.7411	0.1847	0.1645	0.5324
194.	~asón	8	0.002657	0.7844	0.8101	0.3077	0.0458	0.5324
195.	~andolos	44	0.01491	0.7724	0.8097	0.7857	0.7627	0.5324
196.	~irme	23	0.00921	0.7586	0.8287	0.5	0.406	0.5322
197.	~tando	29	0.007105	0.8601	0.7286	0.1768	0.2639	0.5319
198.	~imos	151	0.04843	0.7804	0.7663	0.6681	0.7578	0.5317
199.	~adito	10	0.002928	0.7572	0.8347	0.3333	0.3051	0.5316
200.	~iando	21	0.002285	0.7891	0.8027	0.3088	0.1592	0.5314

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
201.	~taban	20	0.006756	0.862	0.7249	0.1724	0.6434	0.5312
202.	~lados	7	0.02105	0.8672	0.7053	0.07071	0.0468	0.5312
203.	~eta	64	0.01756	0.7019	0.8737	0.4476	0.386	0.5311
204.	~aka	19	0.003208	0.7448	0.8445	0.38	0.7475	0.5308
205.	~isa	118	0.01623	0.7997	0.776	0.6667	0.7014	0.5306
206.	~la	633	0.06791	0.9143	0.6097	0.6349	0.0436	0.5306
207.	~rados	9	0.002721	0.8554	0.7329	0.05696	0.1914	0.5303
208.	~adita	21	0.009832	0.7076	0.8734	0.4667	0.2833	0.5303
209.	~elos	31	0.0235	0.7136	0.8531	0.2897	0.4293	0.5301
210.	~raba	8	0.002868	0.8705	0.716	0.07619	0.07829	0.5298
211.	~ela	48	0.01696	0.7497	0.8221	0.2759	0.5887	0.5296
212.	~ego	7	0.01705	0.8831	0.6884	0.1842	0.02416	0.5295
213.	~isimo	98	0.03038	0.8434	0.714	0.7903	0.7731	0.5292
214.	~etes	20	0.007523	0.6347	0.9454	0.303	0.2458	0.5292
215.	~anse	11	0.002521	0.7885	0.7962	0.2821	0.1563	0.5291
216.	~iados	16	0.002322	0.8265	0.7579	0.2162	0.2377	0.5289
217.	~taba	42	0.007212	0.853	0.7264	0.2456	0.699	0.5289
218.	~edo	7	0.002393	0.7918	0.7921	0.2692	0.141	0.5287
219.	~atibas	32	0.007574	0.8275	0.7508	0.5614	0.7528	0.5286
220.	~ena	18	0.005169	0.75	0.8305	0.2022	0.2487	0.5285
221.	~ono	14	0.002121	0.8311	0.7523	0.3182	0.4108	0.5285
222.	~erían	8	0.0263	0.888	0.6708	0.2667	0.4028	0.5284
223.	~esía	8	0.002903	0.8056	0.7764	0.1569	0.4739	0.5283
224.	~tó	53	0.00542	0.823	0.7564	0.256	0.2945	0.5283
225.	~lada	9	0.004207	0.8626	0.7174	0.08571	0.08352	0.5281
226.	~í	138	0.1031	0.5918	0.8887	0.6635	0.03519	0.5279
227.	~tar	75	0.009633	0.8393	0.734	0.2396	0.3438	0.5276
228.	~ean	25	0.006638	0.7426	0.8337	0.3472	0.7705	0.5276
229.	~isas	22	0.005124	0.7313	0.8463	0.3929	0.3641	0.5276
230.	~radas	12	0.001947	0.8376	0.7425	0.09449	0.1225	0.5273
231.	~irán	52	0.02268	0.7634	0.7955	0.7027	0.553	0.5272
232.	~onas	20	0.006108	0.7684	0.8062	0.3175	0.917	0.5269
233.	~ol	14	0.002322	0.7411	0.8363	0.1414	0.3322	0.5266
234.	~le	613	0.0627	0.9021	0.6132	0.5761	0.2355	0.526
235.	~tan	51	0.008092	0.837	0.7325	0.2099	0.1534	0.5259
236.	~iyas	54	0.02036	0.7607	0.7965	0.4954	0.2446	0.5258
237.	~ese	143	0.1019	0.6025	0.8726	0.6384	0.3907	0.5256
238.	~isima	63	0.02473	0.8302	0.7216	0.7159	0.6093	0.5255
239.	~aña	8	0.002639	0.8312	0.7425	0.2581	0.2474	0.5254
240.	~atas	6	0.001595	0.7375	0.837	0.1111	0.0424	0.5254

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
241.	~tados	35	0.004295	0.8654	0.7058	0.1777	0.2226	0.5252
242.	~eada	21	0.007044	0.7675	0.7994	0.3	0.2607	0.5247
243.	~ías	120	0.01683	0.8562	0.7001	0.4364	0.7187	0.5244
244.	~tara	15	0.005658	0.8554	0.7116	0.2027	0.1724	0.5242
245.	~irlas	8	0.004328	0.7937	0.774	0.3077	0.25	0.524
246.	~iya	107	0.03024	0.7289	0.813	0.6948	0.6119	0.524
247.	~is	115	0.01105	0.795	0.7647	0.27	0.1993	0.5236
248.	~otes	33	0.004572	0.7213	0.8447	0.4459	0.3886	0.5235
249.	~ren	12	0.002557	0.8702	0.6963	0.1165	0.1379	0.523
250.	~eña	15	0.001726	0.7697	0.7968	0.3659	0.204	0.5227
251.	~tada	28	0.003609	0.8772	0.6872	0.1379	0.1653	0.5227
252.	~ían	336	0.1337	0.7244	0.7097	0.7943	0.6899	0.5226
253.	~oneros	8	0.001513	0.905	0.6608	0.4706	0.7018	0.5224
254.	~atibos	30	0.00745	0.8378	0.7205	0.566	0.7034	0.5219
255.	~aderos	6	0.002745	0.7451	0.8169	0.2308	0.05294	0.5216
256.	~rar	21	0.001924	0.8316	0.7307	0.1034	0.09636	0.5214
257.	~itan	6	0.001947	0.873	0.6893	0.15	0.1947	0.5214
258.	~idad	334	0.04171	0.9205	0.6019	0.6326	0.7344	0.5214
259.	~otas	15	0.005358	0.7173	0.8413	0.375	0.5766	0.5213
260.	~tamos	25	0.008011	0.8233	0.7326	0.2451	0.7167	0.5213
261.	~enos	6	0.02106	0.7639	0.7781	0.1	0.7201	0.521
262.	~iga	9	0.003049	0.7834	0.7764	0.2195	0.2783	0.5209
263.	~irla	23	0.01411	0.7396	0.8088	0.4035	0.3217	0.5209
264.	~eka	7	0.003941	0.9014	0.6561	0.14	0.2898	0.5205
265.	~ejo	9	0.00405	0.798	0.759	0.25	0.387	0.5203
266.	~atoria	22	0.003992	0.7794	0.7775	0.55	0.3764	0.5203
267.	~tadas	22	0.002751	0.8436	0.7144	0.1528	0.181	0.5203
268.	~eó	21	0.00439	0.7704	0.7857	0.3134	0.2367	0.5202
269.	~uros	8	0.002164	0.6997	0.8585	0.2162	0.5552	0.5201
270.	~t	36	0.003792	0.6915	0.8648	0.1957	0.1393	0.52
271.	~atorios	8	0.00292	0.8517	0.7038	0.2424	0.1714	0.5195
272.	~elo	60	0.02568	0.7664	0.7662	0.2913	0.3681	0.5194
273.	~iada	14	0.002121	0.8487	0.707	0.1628	0.2126	0.5193
274.	~laban	6	0.01241	0.7683	0.777	0.1714	0.2157	0.5192
275.	~eos	18	0.002385	0.8052	0.7499	0.1895	0.2149	0.5192
276.	~eando	41	0.01003	0.7403	0.8063	0.4316	0.3144	0.5189
277.	~omas	6	0.002322	0.8216	0.7324	0.1765	0.1818	0.5188
278.	~los	410	0.05456	0.9374	0.5627	0.5942	0.02964	0.5182
279.	~eja	12	0.003894	0.7523	0.7983	0.2857	0.4811	0.5182
280.	~ines	36	0.005778	0.7002	0.8481	0.5143	0.7727	0.518

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
281.	~irlos	16	0.01016	0.7323	0.8115	0.3902	0.3284	0.518
282.	~in	11	0.001062	0.6825	0.8703	0.2115	0.0039	0.518
283.	~eres	15	0.01421	0.7231	0.8162	0.2542	0.3725	0.5178
284.	~irá	78	0.03079	0.7545	0.7678	0.8041	0.695	0.5177
285.	~iría	43	0.02208	0.6883	0.8421	0.6719	0.6786	0.5175
286.	~erla	22	0.02113	0.8019	0.7291	0.3667	0.3754	0.5174
287.	~oya	6	0.008351	0.7843	0.7594	0.2727	0.04559	0.5174
288.	~ea	79	0.01682	0.7276	0.8075	0.348	0.2508	0.5173
289.	~las	262	0.03591	0.9253	0.5902	0.4781	0.02841	0.5171
290.	~almente	74	0.00714	0.8091	0.7351	0.4684	0.4342	0.5171
291.	~um	22	0.005387	0.7428	0.8029	0.3235	0.3203	0.517
292.	~onero	10	0.001787	0.7944	0.7542	0.5263	0.8261	0.5168
293.	~esito	20	0.005025	0.6424	0.9021	0.4348	0.2089	0.5165
294.	~gos	8	0.002428	0.8644	0.6818	0.06349	0.04924	0.5162
295.	~esta	13	0.01226	0.8216	0.7145	0.2453	0.01908	0.5161
296.	~r	2587	0.1905	0.9482	0.4096	0.6977	0.5841	0.5161
297.	~osamente	34	0.00551	0.801	0.7415	0.3469	0.412	0.516
298.	~ras	179	0.02089	0.937	0.5893	0.2557	0.41	0.5157
299.	~iyo	75	0.01676	0.7278	0.8024	0.625	0.5117	0.5157
300.	~aje	60	0.008126	0.6925	0.8457	0.566	0.5594	0.5155
301.	~osidad	11	0.001663	0.8001	0.7444	0.2619	0.05085	0.5154
302.	~eko	6	0.01063	0.8481	0.6862	0.2308	0.4861	0.515
303.	~entes	70	0.01229	0.7816	0.75	0.2509	0.3582	0.5146
304.	~inado	7	0.001508	0.7995	0.7427	0.1129	0.08039	0.5146
305.	~iyos	36	0.01112	0.7103	0.8219	0.48	0.2574	0.5144
306.	~ra	917	0.07812	0.8768	0.5882	0.5369	0.6812	0.5144
307.	~ine	6	0.002299	0.7617	0.7788	0.1395	0.2062	0.5143
308.	~esen	16	0.004645	0.7894	0.7479	0.2025	0.1862	0.514
309.	~ansa	23	0.004039	0.7131	0.8242	0.4694	0.4891	0.5138
310.	~tarse	12	0.001947	0.8414	0.6971	0.1008	0.06481	0.5135
311.	~akas	7	0.002352	0.7821	0.7559	0.2333	0.7043	0.5134
312.	~rado	20	0.00273	0.8142	0.7234	0.08511	0.2618	0.5134
313.	~aria	44	0.007021	0.8232	0.7097	0.3056	0.1559	0.5133
314.	~irlo	39	0.02049	0.7133	0.8043	0.629	0.4787	0.5127
315.	~ula	23	0.004987	0.7576	0.7745	0.2987	0.2592	0.5124
316.	~ikamente	63	0.004439	0.8377	0.6947	0.4632	0.3157	0.5123
317.	~rada	14	0.00192	0.8101	0.7245	0.08092	0.1569	0.5122
318.	~ró	12	0.001572	0.8258	0.7089	0.08696	0.07249	0.5121
319.	~erías	24	0.007354	0.7193	0.8089	0.3692	0.3677	0.5118
320.	~iese	9	0.002815	0.8853	0.6472	0.2727	0.256	0.5117

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
321.	~ulas	8	0.003308	0.8233	0.7079	0.16	0.03226	0.5115
322.	~ite	8	0.00197	0.7322	0.8	0.129	0.1266	0.5114
323.	~ele	12	0.02067	0.7998	0.713	0.1463	0.08983	0.5112
324.	~use	9	0.01015	0.8522	0.6709	0.2812	0.6405	0.5111
325.	~tikas	27	0.002591	0.8611	0.6675	0.2061	0.06264	0.5104
326.	~il	42	0.006555	0.6844	0.8402	0.3206	0.1971	0.5104
327.	~imientos	9	0.006208	0.9093	0.6155	0.2	0.1375	0.5103
328.	~udos	8	0.004803	0.7235	0.8024	0.2667	0.3415	0.5102
329.	~lado	14	0.01537	0.8108	0.704	0.1	0.06201	0.51
330.	~esido	10	0.0038	0.7849	0.7411	0.1667	0.2095	0.5099
331.	~itado	8	0.001548	0.8428	0.6853	0.1739	0.258	0.5099
332.	~eado	23	0.00853	0.7646	0.7564	0.3067	0.2869	0.5098
333.	~par	6	0.0202	0.7409	0.7674	0.1622	0.1232	0.5095
334.	~oma	11	0.002943	0.7918	0.7336	0.1897	0.6522	0.5094
335.	~do	2437	0.1749	0.9473	0.4055	0.5597	0.6806	0.5092
336.	~ritas	6	0.001806	0.7858	0.7385	0.1579	0.4755	0.5087
337.	~iles	24	0.004217	0.7222	0.7997	0.2927	0.4061	0.5087
338.	~ates	10	0.002899	0.7438	0.7792	0.25	0.1088	0.5086
339.	~očo	7	0.01783	0.8467	0.6605	0.3684	0.1279	0.5083
340.	~ense	30	0.02566	0.6146	0.8835	0.3571	0.2864	0.5079
341.	~asos	37	0.004219	0.748	0.7706	0.4302	0.6876	0.5076
342.	~abos	7	0.001126	0.7566	0.7636	0.3684	0.1575	0.5071
343.	~tores	47	0.006615	0.8693	0.6448	0.5054	0.6691	0.5069
344.	~tará	18	0.002377	0.7889	0.7292	0.2	0.141	0.5068
345.	~ieran	75	0.0376	0.6849	0.7977	0.7426	0.7848	0.5067
346.	~aro	9	0.003738	0.7501	0.7663	0.225	0.6528	0.5067
347.	~tibos	37	0.01434	0.8943	0.6111	0.296	0.3861	0.5066
348.	~tos	176	0.01231	0.8874	0.619	0.1776	0.3832	0.5062
349.	~tas	254	0.01296	0.89	0.6158	0.2527	0.3896	0.5062
350.	~tarán	7	0.001669	0.8184	0.6985	0.1228	0.1129	0.5062
351.	~iar	37	0.004595	0.7935	0.7205	0.2846	0.25	0.5062
352.	~ta	604	0.02797	0.8497	0.6406	0.351	0.6008	0.5061
353.	~ensias	19	0.006822	0.7418	0.7691	0.181	0.1722	0.5059
354.	~arios	51	0.005784	0.8241	0.6877	0.311	0.3988	0.5059
355.	~ises	17	0.002136	0.8264	0.689	0.3542	0.2804	0.5059
356.	~eño	14	0.002332	0.7104	0.8046	0.2745	0.1348	0.5058
357.	~ene	8	0.01844	0.8626	0.6356	0.1778	0.007287	0.5055
358.	~eme	21	0.0446	0.6438	0.8272	0.3182	0.5636	0.5052
359.	~esita	10	0.004419	0.7126	0.7982	0.303	0.8	0.5051
360.	~ariamos	36	0.02333	0.6296	0.8618	0.878	0.8305	0.5049

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
361.	~ientes	26	0.006214	0.8339	0.6741	0.2796	0.1781	0.5048
362.	~ke	26	0.01449	0.7644	0.7352	0.1757	0.001307	0.5047
363.	~aditas	7	0.004122	0.7456	0.7637	0.25	0.1522	0.5045
364.	~sas	124	0.01249	0.9271	0.5732	0.278	0.1336	0.5043
365.	~tarla	7	0.001709	0.795	0.716	0.1373	0.08333	0.5042
366.	~ote	55	0.01045	0.7283	0.7731	0.3767	0.4739	0.504
367.	~table	7	0.01685	0.7254	0.7684	0.1373	0.2917	0.5036
368.	~alidades	8	0.001636	0.7581	0.7508	0.2667	0.09649	0.5035
369.	~té	20	0.007804	0.76	0.7417	0.2041	0.2706	0.5031
370.	~eas	15	0.004044	0.8103	0.6941	0.1613	0.3458	0.5028
371.	~tika	46	0.004736	0.8409	0.6627	0.1966	0.08758	0.5028
372.	~alisar	6	0.0009618	0.7476	0.759	0.1667	0.04918	0.5025
373.	~istika	9	0.002971	0.7824	0.722	0.2571	0.3803	0.5024
374.	~aduras	11	0.002674	0.6573	0.8472	0.3056	0.3789	0.5024
375.	~erse	105	0.05454	0.7155	0.7359	0.8268	0.8571	0.502
376.	~arias	17	0.005771	0.8182	0.6811	0.191	0.4895	0.5017
377.	~ao	14	0.00562	0.6056	0.8938	0.3256	0.3511	0.5017
378.	~oniko	11	0.001241	0.7279	0.7755	0.2683	0.3459	0.5015
379.	~iste	70	0.03067	0.6098	0.8639	0.6931	0.8919	0.5015
380.	~ua	6	0.003378	0.8806	0.6197	0.08696	0.006037	0.5012
381.	~isadas	12	0.004316	0.8654	0.6319	0.2069	0.1118	0.5005
382.	~ise	25	0.01276	0.7286	0.76	0.4098	0.8487	0.5005
383.	~atibamente	8	0.002498	0.8001	0.6988	0.4	0.7483	0.5004
384.	~iado	31	0.00346	0.7905	0.7073	0.252	0.1402	0.5004
385.	~oče	7	0.01528	0.8553	0.6304	0.5	0.8938	0.5003
386.	~onsito	14	0.004433	0.7037	0.7923	0.56	0.4769	0.5001
387.	~onika	11	0.001471	0.7438	0.755	0.3056	0.2366	0.5001
388.	~ablemente	6	0.002393	0.7583	0.7381	0.1429	0.04204	0.4996
389.	~akos	6	0.00129	0.7634	0.7325	0.2143	0.1966	0.499
390.	~ismos	15	0.01026	0.7596	0.727	0.3	0.7053	0.4989
391.	~tiko	45	0.003678	0.8392	0.6538	0.1957	0.09194	0.4989
392.	~turas	7	0.001046	0.7465	0.7486	0.1228	0.06883	0.4987
393.	~mas	25	0.009633	0.8488	0.6345	0.122	0.06434	0.4976
394.	~alisa	6	0.00244	0.6769	0.8132	0.2	0.06452	0.4975
395.	~andote	14	0.006384	0.71	0.7755	0.5	0.2456	0.4973
396.	~tero	7	0.001086	0.7679	0.7228	0.1045	0.03421	0.4973
397.	~les	455	0.03905	0.9065	0.5463	0.3643	0.3724	0.4973
398.	~ueba	6	0.007624	0.9316	0.5515	0.4615	0.8876	0.4969
399.	~uso	6	0.004129	0.8057	0.6795	0.1935	0.07709	0.4965
400.	~sos	128	0.01007	0.9359	0.5427	0.2876	0.1135	0.4962

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
401.	~so	257	0.01976	0.8962	0.5725	0.3468	0.156	0.4961
402.	~isimas	19	0.008815	0.6948	0.7844	0.5758	0.551	0.496
403.	~ieramos	11	0.006257	0.7891	0.6925	0.3793	0.2105	0.4959
404.	~tamente	17	0.001846	0.7631	0.721	0.2179	0.3195	0.4953
405.	~ama	12	0.01129	0.816	0.6587	0.2143	0.583	0.4953
406.	~enso	6	0.001407	0.8065	0.678	0.2727	0.179	0.4953
407.	~h	18	0.001814	0.673	0.8109	0.2093	0.00952	0.4952
408.	~iano	9	0.0008445	0.7451	0.7395	0.1304	0.3812	0.4951
409.	~isiones	10	0.001675	0.807	0.6766	0.1562	0.5153	0.4951
410.	~ia	204	0.01758	0.8723	0.5945	0.1984	0.3744	0.4948
411.	~toras	10	0.00235	0.7967	0.685	0.4348	0.5147	0.4947
412.	~eser	10	0.01084	0.7529	0.7178	0.1667	0.3305	0.4938
413.	~sa	281	0.01933	0.8792	0.5822	0.3168	0.249	0.4936
414.	~ario	76	0.006102	0.7652	0.7092	0.3619	0.3036	0.4935
415.	~ín	55	0.007595	0.6232	0.8484	0.5189	0.5333	0.4931
416.	~itar	7	0.001991	0.7728	0.7043	0.1296	0.4443	0.493
417.	~roso	7	0.001186	0.8637	0.6142	0.1111	0.1508	0.493
418.	~añas	8	0.001988	0.7755	0.7012	0.2857	0.1237	0.4929
419.	~enses	6	0.001032	0.7443	0.7333	0.2	0.2333	0.4929
420.	~re	60	0.004204	0.7331	0.7412	0.1807	0.3101	0.4928
421.	~isimos	33	0.01493	0.7334	0.729	0.6735	0.4427	0.4924
422.	~riko	7	0.001367	0.8621	0.6131	0.06306	0.03158	0.4922
423.	~san	11	0.001663	0.8755	0.5991	0.07639	0.04629	0.4921
424.	~ansia	22	0.004536	0.7307	0.7397	0.3099	0.3359	0.4916
425.	~enes	10	0.002252	0.7085	0.7637	0.2041	0.2124	0.4915
426.	~men	6	0.02601	0.8173	0.6307	0.1	0.1574	0.4914
427.	~nas	77	0.01062	0.9047	0.5585	0.1812	0.2427	0.4913
428.	~lote	6	0.01232	0.752	0.7089	0.24	0.4138	0.4911
429.	~oh	6	0.01278	0.7019	0.7577	0.4286	0.06383	0.4908
430.	~tido	7	0.002091	0.827	0.6429	0.1186	0.01548	0.4907
431.	~andonos	18	0.01055	0.6514	0.8092	0.5294	0.5122	0.4904
432.	~isar	69	0.008135	0.8319	0.6308	0.4964	0.3195	0.4903
433.	~alisasión	18	0.0033	0.7759	0.6915	0.3396	0.2705	0.4903
434.	~ie	13	0.002154	0.7377	0.7309	0.1857	0.02417	0.4902
435.	~tarlo	8	0.001179	0.7549	0.7134	0.125	0.07812	0.4898
436.	~gas	10	0.01068	0.7322	0.7262	0.09615	0.07626	0.4897
437.	~ses	33	0.00569	0.8162	0.6466	0.1398	0.1128	0.4895
438.	~ere	12	0.01633	0.8125	0.638	0.1875	0.1263	0.4889
439.	~tiba	54	0.01783	0.8432	0.6055	0.3354	0.316	0.4888
440.	~emos	360	0.1474	0.6496	0.6693	0.8978	0.5812	0.4888

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
441.	~lan	9	0.001486	0.7101	0.7543	0.1023	0.2986	0.4886
442.	~edor	12	0.01423	0.707	0.7438	0.2791	0.6466	0.4883
443.	~to	366	0.02257	0.8365	0.6048	0.215	0.4417	0.4879
444.	~ares	46	0.006636	0.8371	0.6196	0.3046	0.4348	0.4878
445.	~día	8	0.01246	0.8569	0.5929	0.1053	0.03785	0.4874
446.	~ran	141	0.0345	0.8891	0.5383	0.2975	0.2034	0.4873
447.	~isan	16	0.004565	0.7927	0.6641	0.2759	0.2188	0.4871
448.	~rosa	9	0.001126	0.7807	0.6793	0.1579	0.3299	0.487
449.	~alismo	22	0.002469	0.7157	0.7427	0.3929	0.4129	0.4869
450.	~tibo	63	0.01908	0.8655	0.5759	0.3663	0.3349	0.4868
451.	~ernos	11	0.01338	0.7167	0.7304	0.22	0.1862	0.4868
452.	~tibas	39	0.01536	0.8933	0.5511	0.312	0.2724	0.4866
453.	~idades	86	0.01412	0.8402	0.6053	0.6772	0.5826	0.4865
454.	~ios	64	0.00581	0.8863	0.5673	0.1382	0.1765	0.4865
455.	~uela	16	0.01916	0.6783	0.7612	0.5	0.8308	0.4862
456.	~isado	35	0.009639	0.8281	0.6207	0.3431	0.2031	0.4861
457.	~idores	11	0.004529	0.7107	0.7423	0.5	0.4327	0.4858
458.	~ajo	16	0.007187	0.6982	0.7517	0.3478	0.2887	0.4857
459.	~ago	8	0.01089	0.6667	0.7784	0.2667	0.07573	0.4853
460.	~enta	18	0.004066	0.7434	0.7083	0.1651	0.12	0.4852
461.	~bar	10	0.002519	0.7543	0.6983	0.1111	0.2224	0.485
462.	~iendose	68	0.03534	0.6995	0.7201	0.701	0.6156	0.485
463.	~taría	7	0.001146	0.7837	0.6695	0.1373	0.04583	0.4848
464.	~ien	7	0.00197	0.7654	0.6866	0.2059	0.9835	0.4847
465.	~pe	9	0.00527	0.7091	0.7393	0.1731	0.07289	0.4845
466.	~ben	9	0.01484	0.7729	0.6649	0.1324	0.2692	0.4842
467.	~ias	60	0.006076	0.9093	0.5373	0.1345	0.2211	0.4842
468.	~c	11	0.00174	0.6206	0.8301	0.131	0.02034	0.4842
469.	~tora	7	0.004283	0.8233	0.6234	0.1892	0.08434	0.4837
470.	~sita	18	0.00918	0.783	0.6582	0.1837	0.09266	0.4834
471.	~tor	51	0.005889	0.8556	0.588	0.573	0.6547	0.4832
472.	~enado	6	0.002838	0.6696	0.7757	0.1622	0.3562	0.4827
473.	~nan	6	0.001103	0.8668	0.5793	0.04762	0.01437	0.4824
474.	~erlos	15	0.02167	0.7302	0.6951	0.3	0.2892	0.4823
475.	~ramos	31	0.005725	0.8337	0.6074	0.2313	0.2597	0.4823
476.	~iente	51	0.0108	0.7125	0.723	0.3835	0.1987	0.4821
477.	~iere	11	0.003365	0.7199	0.7226	0.3333	0.3239	0.482
478.	~ide	6	0.001736	0.7893	0.6548	0.1154	0.4247	0.4819
479.	~elas	20	0.01203	0.6719	0.7617	0.1818	0.5137	0.4819
480.	~mos	1103	0.182	0.857	0.4064	0.6911	0.6531	0.4818

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
481.	~erlas	6	0.01178	0.7746	0.6584	0.1714	0.3368	0.4816
482.	~irle	12	0.01429	0.6264	0.8023	0.3429	0.07234	0.481
483.	~onada	6	0.004387	0.9601	0.4774	0.1071	0.0402	0.4806
484.	~isó	15	0.004832	0.8229	0.6139	0.2679	0.1161	0.4805
485.	~idamente	9	0.005098	0.6497	0.7853	0.2571	0.06494	0.4801
486.	~aramos	26	0.01475	0.5782	0.8471	0.6842	0.7143	0.48
487.	~tikos	34	0.001933	0.7876	0.6505	0.1965	0.1295	0.48
488.	~gó	7	0.002855	0.7429	0.694	0.09333	0.03234	0.4799
489.	~ora	146	0.0115	0.908	0.5185	0.4465	0.1042	0.4793
490.	~el	28	0.009204	0.5259	0.9019	0.2718	0.008008	0.479
491.	~na	185	0.01861	0.8714	0.5458	0.1943	0.1398	0.4786
492.	~isión	39	0.004854	0.7002	0.7296	0.3023	0.3364	0.4782
493.	~io	182	0.01657	0.7897	0.6277	0.2476	0.3965	0.478
494.	~ones	815	0.04329	0.9561	0.4336	0.8463	0.9221	0.4777
495.	~alidad	50	0.0063	0.7045	0.7205	0.4762	0.385	0.4771
496.	~te	1072	0.08962	0.8838	0.4571	0.3589	0.4897	0.4769
497.	~só	6	0.001337	0.8332	0.5948	0.04511	0.03141	0.4764
498.	~tra	12	0.003014	0.8222	0.6035	0.1791	0.009532	0.4762
499.	~ensia	83	0.0176	0.6118	0.7988	0.343	0.534	0.4761
500.	~sitos	28	0.003167	0.8399	0.5848	0.3218	0.1233	0.476
501.	~are	9	0.002956	0.6627	0.7619	0.36	0.3846	0.4759
502.	~iales	10	0.002125	0.8207	0.6039	0.1031	0.1644	0.4756
503.	~imiento	83	0.02182	0.7204	0.683	0.7217	0.6658	0.4751
504.	~me	565	0.08237	0.8561	0.4848	0.7869	0.1243	0.4744
505.	~gado	9	0.002377	0.7234	0.6972	0.1	0.07879	0.4743
506.	~m	21	0.00319	0.7089	0.7106	0.1364	0.1066	0.4742
507.	~adura	11	0.002789	0.5984	0.8205	0.3438	0.3448	0.4739
508.	~iadas	10	0.002716	0.7076	0.7104	0.1639	0.1032	0.4736
509.	~da	441	0.1554	0.9297	0.3343	0.1996	0.3044	0.4732
510.	~erlo	35	0.03475	0.6804	0.7036	0.493	0.2995	0.4729
511.	~mente	981	0.08863	0.9758	0.3539	0.8814	0.9763	0.4728
512.	~tasión	23	0.001536	0.6996	0.7171	0.1474	0.1925	0.4727
513.	~aderas	6	0.001595	0.632	0.7845	0.3333	0.1368	0.4727
514.	~sen	31	0.009548	0.8241	0.5834	0.1582	0.09403	0.4723
515.	~saba	7	0.001126	0.8008	0.6151	0.07216	0.06004	0.4723
516.	~teros	6	0.0009852	0.7066	0.7088	0.12	0.07273	0.4721
517.	~iaba	9	0.001939	0.6901	0.7235	0.1837	0.225	0.4718
518.	~ma	45	0.008736	0.7818	0.6248	0.1169	0.05749	0.4718
519.	~onado	8	0.005348	0.9278	0.481	0.1231	0.03056	0.4714
520.	~das	174	0.1732	0.9294	0.3114	0.1162	0.09711	0.4713

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
521.	~ibles	13	0.001776	0.7036	0.7085	0.1461	0.07824	0.4713
522.	~res	334	0.02396	0.9553	0.434	0.4123	0.3515	0.4711
523.	~iré	14	0.009189	0.6873	0.7164	0.4	0.09375	0.471
524.	~iendolo	12	0.01219	0.6837	0.717	0.3333	0.2787	0.4709
525.	~li	7	0.009068	0.6786	0.7242	0.1458	0.1778	0.4706
526.	~ular	9	0.002502	0.7931	0.6156	0.08654	0.1122	0.4704
527.	~istiko	12	0.002475	0.696	0.7089	0.375	0.4	0.4691
528.	~nar	6	0.001712	0.8481	0.5567	0.0339	0.01503	0.4688
529.	~oles	14	0.01265	0.7364	0.6561	0.175	0.4	0.4684
530.	~onadas	8	0.002709	0.8485	0.5507	0.2	0.09649	0.4673
531.	~eles	11	0.005681	0.6461	0.75	0.1833	0.06093	0.4673
532.	~ian	11	0.001881	0.7497	0.6472	0.193	0.2056	0.4662
533.	~erte	18	0.02327	0.6265	0.7483	0.2903	0.2872	0.466
534.	~no	108	0.01043	0.8228	0.5645	0.1785	0.03952	0.4659
535.	~esitos	7	0.004484	0.6026	0.7902	0.2414	0.2308	0.4658
536.	~ís	21	0.002654	0.6439	0.7497	0.3281	0.204	0.4654
537.	~ible	31	0.005875	0.6846	0.705	0.2605	0.1148	0.4652
538.	~isada	23	0.006138	0.7948	0.5943	0.2771	0.1604	0.4651
539.	~oras	26	0.003957	0.883	0.5077	0.1667	0.04167	0.4649
540.	~esió	8	0.002745	0.7762	0.6149	0.1509	0.1477	0.4646
541.	~nos	361	0.04589	0.8013	0.5458	0.5187	0.256	0.4643
542.	~jas	7	0.002815	0.8224	0.5673	0.0814	0.1056	0.4642
543.	~ós	8	0.0008621	0.6587	0.7324	0.4444	0.2252	0.464
544.	~sía	12	0.001724	0.8462	0.5437	0.1176	0.09925	0.4639
545.	~lar	18	0.003722	0.777	0.611	0.07692	0.06169	0.4639
546.	~gar	9	0.005364	0.6942	0.6919	0.09375	0.01858	0.4638
547.	~andoles	13	0.008856	0.6069	0.7754	0.5652	0.3409	0.4637
548.	~eaba	13	0.003443	0.5664	0.8209	0.2766	0.1462	0.4636
549.	~tibamente	10	0.006559	0.7899	0.5937	0.2174	0.2568	0.4634
550.	~uras	24	0.01247	0.8856	0.4917	0.1832	0.1146	0.4632
551.	~sito	64	0.00559	0.7907	0.5923	0.4638	0.2993	0.4629
552.	~ifikada	7	0.002594	0.7431	0.6404	0.1944	0.1026	0.462
553.	~dos	214	0.165	0.92	0.2998	0.1179	0.2397	0.4616
554.	~go	56	0.01356	0.6763	0.6949	0.2213	0.2453	0.4616
555.	~ai	8	0.002182	0.594	0.7876	0.25	0.01028	0.4613
556.	~ro	155	0.009523	0.8554	0.5189	0.2089	0.09327	0.4613
557.	~be	32	0.01001	0.7179	0.6553	0.2238	0.353	0.4611
558.	~uye	6	0.002463	0.7304	0.6494	0.1714	0.1027	0.4608
559.	~eaban	12	0.004621	0.6012	0.7756	0.2927	0.2154	0.4605
560.	~íos	7	0.004504	0.6857	0.6909	0.2121	0.2903	0.4604

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
561.	~de	34	0.007799	0.7669	0.6046	0.1683	0.004353	0.4598
562.	~atorias	8	0.002516	0.6882	0.6883	0.4	0.1385	0.4597
563.	~nales	6	0.001337	0.8932	0.4838	0.04878	0.1609	0.4594
564.	~rías	23	0.006413	0.8491	0.5223	0.1901	0.3238	0.4593
565.	~ga	42	0.01116	0.6661	0.6996	0.1803	0.08511	0.459
566.	~lasión	6	0.001337	0.7797	0.5945	0.05505	0.01368	0.4585
567.	~ores	189	0.01575	0.9151	0.4443	0.3803	0.3472	0.4584
568.	~tes	183	0.01844	0.9163	0.4398	0.2047	0.1812	0.4582
569.	~isando	9	0.003894	0.7592	0.6101	0.18	0.08763	0.4577
570.	~ura	88	0.01765	0.7607	0.5947	0.4112	0.5254	0.4577
571.	~irnos	18	0.012	0.5818	0.7789	0.4615	0.3226	0.4576
572.	~esko	7	0.001769	0.7502	0.6205	0.2059	0.05116	0.4575
573.	~or	324	0.0202	0.8663	0.4853	0.485	0.2685	0.4573
574.	~nada	6	0.001337	0.7881	0.5822	0.03488	0.002891	0.4572
575.	~rito	9	0.001314	0.683	0.6839	0.1452	0.1903	0.4561
576.	~arselo	10	0.007431	0.6232	0.7353	0.4545	0.4615	0.4553
577.	~dió	12	0.003882	0.8332	0.5279	0.1644	0.1512	0.455
578.	~tibidad	10	0.003336	0.7905	0.57	0.3125	0.2803	0.4546
579.	~oide	9	0.001517	0.6499	0.7122	0.4286	0.5	0.4545
580.	~isia	9	0.003362	0.6666	0.6931	0.2432	0.14	0.4544
581.	~one	12	0.000821	0.8281	0.5316	0.2143	0.0274	0.4535
582.	~tura	19	0.001882	0.6659	0.6915	0.1979	0.1168	0.4531
583.	~tito	8	0.001495	0.7301	0.6267	0.1311	0.07884	0.4528
584.	~sando	6	0.001337	0.7716	0.5852	0.05	0.03542	0.4527
585.	~ría	405	0.06003	0.7916	0.5064	0.6841	0.4654	0.4527
586.	~isados	17	0.006276	0.7855	0.5645	0.2208	0.1081	0.4521
587.	~alista	13	0.002468	0.6383	0.7152	0.2453	0.188	0.452
588.	~jos	10	0.002562	0.8198	0.5318	0.1299	0.2447	0.4514
589.	~ja	23	0.009669	0.729	0.6139	0.1565	0.1726	0.4509
590.	~ne	32	0.00289	0.7288	0.6206	0.1649	0.04233	0.4508
591.	~arías	6	0.01088	0.5295	0.811	0.2857	0.175	0.4505
592.	~če	11	0.009123	0.6625	0.6779	0.1486	0.1035	0.4498
593.	~po	11	0.008586	0.6273	0.7124	0.1358	0.0682	0.4494
594.	~irían	11	0.01308	0.5493	0.7838	0.3143	0.2295	0.4487
595.	~isasión	106	0.009848	0.8608	0.4749	0.6625	0.4466	0.4485
596.	~atorio	13	0.005034	0.6325	0.7057	0.325	0.1755	0.4477
597.	~diendo	6	0.004105	0.8276	0.5102	0.1111	0.04971	0.4473
598.	~selo	7	0.01359	0.8191	0.5088	0.1148	0.07692	0.4472
599.	~ča	11	0.002418	0.6723	0.6664	0.1122	0.006272	0.447
600.	~iana	6	0.001337	0.6584	0.6802	0.1071	0.1481	0.4466

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
601.	~lón	8	0.004609	0.6834	0.6518	0.1905	0.05163	0.4466
602.	~ulares	10	0.001619	0.7088	0.6283	0.1724	0.1371	0.4463
603.	~osis	12	0.001067	0.6197	0.7156	0.1579	0.425	0.4454
604.	~dieron	8	0.003958	0.8109	0.5203	0.1429	0.04158	0.445
605.	~gan	10	0.004645	0.6784	0.6519	0.1042	0.03783	0.445
606.	~olojía	18	0.00344	0.5956	0.7358	0.2647	0.209	0.445
607.	~uales	9	0.002971	0.6875	0.6442	0.2368	0.5382	0.4449
608.	~abilidad	19	0.006022	0.6308	0.6973	0.3958	0.734	0.4447
609.	~ge	8	0.002551	0.6735	0.6579	0.1481	0.05	0.4447
610.	~ros	27	0.005984	0.86	0.4666	0.05305	0.02514	0.4442
611.	~rán	204	0.06982	0.8623	0.3971	0.4647	0.2881	0.4431
612.	~ón	957	0.05114	0.9197	0.3581	0.4673	0.7699	0.443
613.	~niko	17	0.00202	0.7803	0.5463	0.1574	0.1585	0.4429
614.	~iremos	9	0.01134	0.6445	0.6721	0.3103	0.1028	0.4426
615.	~jo	20	0.003941	0.7393	0.5832	0.119	0.01492	0.4422
616.	~atura	9	0.001548	0.5314	0.7912	0.3462	0.1088	0.4414
617.	~x	10	0.001548	0.5273	0.7943	0.1493	0.06538	0.441
618.	~ié	8	0.001654	0.6239	0.6976	0.2286	0.2015	0.441
619.	~tones	10	0.0007882	0.6905	0.6278	0.2703	0.1111	0.4397
620.	~ifikasión	23	0.009577	0.6856	0.6232	0.3833	0.2528	0.4395
621.	~nado	18	0.002541	0.7312	0.5837	0.08333	0.1351	0.4391
622.	~sado	14	0.003006	0.7394	0.5736	0.06222	0.03927	0.4387
623.	~ulo	15	0.002918	0.7356	0.5738	0.1875	0.2588	0.4374
624.	~ré	66	0.04632	0.7958	0.4689	0.2773	0.1718	0.437
625.	~ual	12	0.004363	0.6822	0.6221	0.2609	0.4954	0.4362
626.	~tón	12	0.001712	0.7237	0.5783	0.2222	0.06826	0.4346
627.	~des	29	0.00249	0.8015	0.4961	0.09006	0.2086	0.4334
628.	~ioso	8	0.002076	0.6745	0.6231	0.1176	0.04115	0.4332
629.	~eska	10	0.003561	0.6741	0.6203	0.2	0.06667	0.4326
630.	~emia	6	0.0006568	0.606	0.6906	0.2	0.08537	0.4324
631.	~l	358	0.01792	0.8039	0.4743	0.2775	0.08624	0.4321
632.	~iamos	96	0.04253	0.5778	0.6749	0.6358	0.4386	0.4318
633.	~ria	18	0.009172	0.8489	0.437	0.06618	0.1206	0.4317
634.	~rte	83	0.05729	0.8627	0.3746	0.3051	0.1091	0.4315
635.	~tismo	10	0.001239	0.5871	0.7019	0.1852	0.2827	0.4301
636.	~amento	15	0.003847	0.5042	0.7796	0.4412	0.3678	0.4292
637.	~ienta	6	0.002205	0.6821	0.6017	0.1875	0.2905	0.4287
638.	~én	9	0.001454	0.5677	0.7164	0.2093	0.7762	0.4285
639.	~sio	11	0.001088	0.681	0.5993	0.1038	0.1724	0.4271
640.	~ial	32	0.004214	0.7157	0.5606	0.2078	0.3879	0.4268

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
641.	~sidad	28	0.003036	0.7754	0.5019	0.2667	0.4075	0.4268
642.	~yo	38	0.002985	0.8457	0.4296	0.19	0.0147	0.4261
643.	~kan	7	0.002352	0.7161	0.5595	0.05556	0.06826	0.426
644.	~čes	6	0.01173	0.6185	0.6445	0.1463	0.1527	0.4249
645.	~erme	48	0.03376	0.5238	0.7165	0.6957	0.6434	0.4247
646.	~rse	786	0.1131	0.9031	0.2568	0.7429	0.7201	0.4243
647.	~ío	16	0.004416	0.5634	0.7017	0.2424	0.04863	0.4232
648.	~rle	56	0.07061	0.9243	0.2735	0.2171	0.1266	0.4228
649.	~siones	324	0.03644	0.7195	0.5112	0.48	0.4065	0.4224
650.	~rá	462	0.08029	0.836	0.3499	0.7276	0.42	0.4221
651.	~iba	15	0.009073	0.822	0.435	0.06122	0.01728	0.422
652.	~mo	118	0.004128	0.8394	0.4221	0.1835	0.03276	0.4219
653.	~tario	6	0.0008679	0.6853	0.5779	0.1071	0.06238	0.4214
654.	~sión	863	0.0405	0.7837	0.4398	0.5575	0.5586	0.4213
655.	~edad	24	0.01371	0.6169	0.6327	0.5	0.361	0.4211
656.	~bo	15	0.004616	0.7997	0.4519	0.04274	0.00948	0.4187
657.	~rían	76	0.02813	0.7597	0.4671	0.4368	0.1972	0.4183
658.	~sis	45	0.003297	0.6899	0.5614	0.2663	0.4929	0.4182
659.	~rás	32	0.018	0.7412	0.4936	0.2832	0.08485	0.4176
660.	~itis	9	0.001267	0.5366	0.7146	0.1875	0.2772	0.4175
661.	~torio	10	0.004405	0.8032	0.4441	0.1515	0.1459	0.4172
662.	~ya	62	0.005505	0.7636	0.4799	0.2199	0.02945	0.4163
663.	~remos	62	0.03572	0.7672	0.4439	0.337	0.2416	0.4156
664.	~ifika	12	0.009383	0.604	0.6331	0.24	0.09263	0.4155
665.	~erá	68	0.04209	0.5923	0.6114	0.8395	0.9594	0.4153
666.	~rme	202	0.06024	0.8456	0.3373	0.5088	0.4337	0.4144
667.	~rlo	140	0.08196	0.8873	0.2692	0.2923	0.1588	0.4128
668.	~ola	13	0.00838	0.6907	0.5379	0.08176	0.09007	0.4123
669.	~'s	13	0.001029	0.5071	0.7285	0.4333	0.09187	0.4122
670.	~ndo	345	0.09178	0.9018	0.2397	0.2189	0.1865	0.4111
671.	~ás	36	0.008965	0.6985	0.5205	0.2045	0.1543	0.4093
672.	~ron	317	0.0828	0.9387	0.2056	0.2805	0.4459	0.409
673.	~rio	35	0.007307	0.8169	0.3977	0.1029	0.0681	0.4073
674.	~sina	7	0.002473	0.5942	0.6234	0.125	0.2514	0.4067
675.	~tro	7	0.001468	0.6248	0.5925	0.07143	0.00161	0.4062
676.	~čo	7	0.003117	0.5703	0.6396	0.07447	0.00298	0.4043
677.	~csión	6	0.002463	0.7341	0.4761	0.04762	0.01746	0.4042
678.	~nal	10	0.002196	0.7545	0.4545	0.0641	0.2237	0.4037
679.	~iensia	6	0.000563	0.7321	0.4745	0.1875	0.1158	0.4024
680.	~dón	6	0.0005161	0.5985	0.608	0.2222	0.4737	0.4023

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
681.	~ral	10	0.001393	0.6033	0.6001	0.08929	0.3764	0.4016
682.	~rlas	18	0.04007	0.8765	0.2827	0.08219	0.04945	0.3998
683.	~isarse	7	0.00382	0.6307	0.5621	0.1321	0.06299	0.3989
684.	~ndole	14	0.0189	0.9447	0.2329	0.1176	0.05993	0.3988
685.	~ndose	43	0.04557	0.9532	0.1969	0.1083	0.09568	0.3986
686.	~riamos	12	0.02194	0.7394	0.4318	0.1765	0.1095	0.3977
687.	~ba	148	0.1046	0.7718	0.3138	0.1146	0.0438	0.3967
688.	~je	44	0.004725	0.6465	0.5369	0.2245	0.1876	0.396
689.	~sko	8	0.001179	0.7005	0.4864	0.1053	0.2817	0.396
690.	~erán	21	0.02411	0.6137	0.5501	0.4038	0.07029	0.396
691.	~sar	16	0.002648	0.671	0.5126	0.06275	0.03485	0.3954
692.	~nse	7	0.007279	0.8623	0.315	0.05224	0.01178	0.3949
693.	~ús	7	0.0004826	0.4939	0.6896	0.3684	0.07255	0.3947
694.	~rnos	115	0.04633	0.7766	0.3567	0.4492	0.2494	0.3932
695.	~rla	93	0.05533	0.8383	0.2841	0.2268	0.1025	0.3926
696.	~isidad	11	0.001331	0.5078	0.6647	0.3235	0.227	0.3913
697.	~emente	7	0.00185	0.7305	0.441	0.04192	0.0177	0.3911
698.	~dor	108	0.08117	0.8437	0.2459	0.2512	0.1775	0.3902
699.	~ei	6	0.007014	0.3301	0.8325	0.2069	0.03987	0.3899
700.	~rs	13	0.001581	0.6586	0.5081	0.2955	0.3276	0.3894
701.	~ridad	9	0.001439	0.6292	0.5275	0.1731	0.2998	0.3861
702.	~rios	9	0.00699	0.7816	0.3668	0.03571	0.005946	0.3851
703.	~ste	118	0.06248	0.7075	0.3816	0.3894	0.06129	0.3839
704.	~cto	6	0.001103	0.6587	0.4897	0.06897	0.01134	0.3832
705.	~dero	11	0.004952	0.6817	0.4628	0.1642	0.1309	0.3831
706.	~rles	14	0.02482	0.81	0.3124	0.1167	0.05782	0.3824
707.	~endo	14	0.01091	0.8747	0.2518	0.03804	0.05332	0.3792
708.	~nidad	7	0.0009651	0.5175	0.6145	0.1489	0.04534	0.3776
709.	~án	52	0.02411	0.794	0.3065	0.09683	0.08989	0.3749
710.	~ko	90	0.01522	0.7682	0.3386	0.08523	0.1409	0.374
711.	~itud	14	0.003549	0.5628	0.5553	0.4828	0.5	0.3739
712.	~dores	77	0.06789	0.8293	0.2235	0.2476	0.1599	0.3736
713.	~ente	84	0.01304	0.743	0.3602	0.05722	0.1314	0.3721
714.	~d	48	0.006636	0.817	0.2916	0.06138	0.2854	0.3718
715.	~rlos	42	0.04115	0.8005	0.2717	0.1346	0.08362	0.3711
716.	~ka	78	0.02002	0.7109	0.3736	0.07379	0.1317	0.3682
717.	~nte	212	0.02805	0.8639	0.2058	0.1129	0.1429	0.3659
718.	~sia	154	0.004068	0.7166	0.3743	0.3415	0.2281	0.365
719.	~ño	6	0.001196	0.651	0.4308	0.06122	0.007879	0.361
720.	~ntes	40	0.01869	0.7838	0.2747	0.06838	0.008026	0.3591

Tabla C.1 (continuación):
Sufijos del *CEMC* en orden de *afijalidad*

	afijo	fr.	cdrs.	econ.	entrop.	prob1	prob2	afijalidad
721.	~sta	11	0.002354	0.7925	0.2807	0.02523	0.005392	0.3585
722.	~ble	87	0.02648	0.7529	0.2913	0.218	0.3324	0.3569
723.	~nsia	32	0.006219	0.8298	0.2228	0.09969	0.04134	0.3529
724.	~eron	11	0.01332	0.8468	0.1963	0.03642	0.01192	0.3521
725.	~miento	248	0.03863	0.7726	0.2388	0.7086	0.704	0.35
726.	~kos	20	0.00651	0.7002	0.3264	0.02933	0.007924	0.3444
727.	~nes	31	0.01011	0.7908	0.2272	0.02703	0.006139	0.3427
728.	~yas	6	0.002838	0.6385	0.3867	0.03529	0.004904	0.3427
729.	~mento	7	0.004524	0.4651	0.5539	0.08974	0.09004	0.3412
730.	~'l	9	0.001407	0.6133	0.4087	0.4091	0.2273	0.3412
731.	~dora	32	0.05052	0.7567	0.215	0.1538	0.104	0.3407
732.	~ña	6	0.003472	0.5805	0.4372	0.06452	0.01802	0.3404
733.	~á	70	0.02876	0.6968	0.2933	0.09091	0.2067	0.3396
734.	~on	65	0.03098	0.8257	0.155	0.05263	0.003498	0.3372
735.	~kas	14	0.008203	0.6392	0.364	0.02422	0.005862	0.3371
736.	~nta	7	0.002091	0.6181	0.3808	0.04046	0.05199	0.3336
737.	~ska	8	0.001003	0.5502	0.4327	0.08989	0.06499	0.328
738.	~iento	8	0.005454	0.7966	0.1624	0.02089	0.002313	0.3215
739.	~lidad	8	0.007354	0.6106	0.3091	0.0362	0.04913	0.309
740.	~mientos	12	0.01135	0.6529	0.2562	0.1034	0.01907	0.3068
741.	~bles	16	0.02329	0.6073	0.2614	0.05369	0.0135	0.2973
742.	~bilidad	7	0.002272	0.5411	0.3437	0.08046	0.04526	0.2957
743.	~ión	41	0.02525	0.7454	0.1107	0.02512	0.03188	0.2938
744.	~dad	18	0.004066	0.6622	0.1815	0.0298	0.08564	0.2826
745.	~tis	7	0.0004826	0.4956	0.3213	0.1148	0.1141	0.2725
746.	~ban	22	0.04616	0.588	0.1481	0.03313	0.007404	0.2607
747.	~ad	22	0.004721	0.6438	0.1065	0.03438	0.1334	0.2517
748.	~ilidad	8	0.001707	0.4394	0.2538	0.07547	0.01245	0.2316
749.	~smo	6	0.006803	0.4747	0.06093	0.01644	0.004553	0.1808

Tabla C.2: Algunas agrupaciones por forma de sufijos derivativos observadas en la tabla C.1 que no se examinaron en el capítulo sobre el afijo.

sufijos	fr.	afijalidad
~ismo	213	0.5632
~ismos	15	0.4989
~alismo	22	0.4869
~isima	63	0.5255
~isimas	19	0.496
~isimo	98	0.5292
~isimos	33	0.4924
~ota	24	0.5375
~otas	15	0.5213
~ote	55	0.504
~otes	33	0.5235
~uro	16	0.539
~uros	8	0.5201
~ista	231	0.5555
~istas	181	0.5651
~alista	13	0.452
~osis	12	0.4454
~sis	45	0.4182
~itis	9	0.4175
~tis	7	0.2725
~oide	9	0.4545
~emia	6	0.4324
~oma	11	0.5094
~omas	6	0.5188
~udo	21	0.549
~udos	8	0.5102
~ular	9	0.4704
~ulares	10	0.4463
~lar	18	0.4639
~anta	16	0.5733
~enta	18	0.4852
~iento	8	0.3215
~ienta	6	0.4287
~nta	7	0.3336
~ulo	15	0.4374
~ula	23	0.5124
~ulas	8	0.5115
~uela	16	0.4862

Tabla C.2 (continuación): Algunas agrupaciones por forma de sufijos derivativos observadas en la tabla C.1 que no se examinaron en el capítulo sobre el afijo.

sufijos	fr.	afijalidad
~án	52	0.3749
~anes	29	0.558
~ano	67	0.5493
~anos	53	0.5396
~ana	58	0.5686
~anas	18	0.5577
~iana	6	0.4466
~iano	9	0.4951
~eña	15	0.5227
~eño	14	0.5058
~enses	6	0.4929
~és	89	0.5357
~eses	26	0.5613
~esa	51	0.5368
~esas	24	0.5551
~ol	14	0.5266
~ola	13	0.4123
~oles	14	0.4684
~año	9	0.5416
~años	7	0.5857
~aña	8	0.5254
~añas	8	0.4929
~ato	60	0.5369
~ata	32	0.5653
~atos	20	0.5365
~atas	6	0.5254
~ates	10	0.5086
~eto	10	0.5626
~etos	6	0.5464
~eta	64	0.5311
~etas	31	0.5399
~etes	20	0.5292
~ite	8	0.5114

Tabla C.2 (continuación):
Algunas agrupaciones por forma de sufijos derivativos observadas en la tabla C.1 que no se examinaron en el capítulo sobre el afijo.

sufijos	fr.	afijalidad
~aje	60	0.5155
~ajes	20	0.534
~ajo	16	0.4857
~eja	12	0.5182
~ejo	9	0.5203
~ija	7	0.5682
~ijo	10	0.5669
~oja	7	0.554
~je	44	0.396
~ako	10	0.5451
~akos	6	0.499
~eka	7	0.5205
~eko	6	0.515
~eska	10	0.4326
~esko	7	0.4575
~sko	8	0.396
~el	28	0.479
~ela	48	0.5296
~elo	60	0.5194
~il	42	0.5104
~ila	16	0.5452
~ilado	6	0.5612
~iles	24	0.5087
~ilo	14	0.5764
~én	9	0.4285
~ena	18	0.5285
~enas	10	0.5357
~eno	15	0.5413
~enos	6	0.521
~ene	8	0.5055
~enes	10	0.4915
~ono	14	0.5285

Tabla C.2 (continuación):
Algunas agrupaciones por forma de sufijos derivativos observadas en la tabla C.1 que no se examinaron en el capítulo sobre el afijo.

sufijos	fr.	afijalidad
~ín	55	0.4931
~ina	115	0.5483
~inado	7	0.5146
~inas	38	0.5498
~ines	36	0.518
~ino	48	0.55
~inos	34	0.5356
~aso	94	0.5496
~asa	21	0.558
~asas	9	0.5973
~asos	37	0.5076
~ases	19	0.5664
~asón	8	0.5324
~eso	17	0.5828
~esos	8	0.5672
~és	89	0.5357
~eses	26	0.5613
~is	115	0.5236
~ís	21	0.4654
~ises	17	0.5059
~ise	25	0.5005
~iso	28	0.5553
~isos	11	0.5405
~ós	8	0.464
~aya	13	0.558
~ayo	9	0.5681
~oya	6	0.5174

Apéndice D

Las formas más gramaticales del *CEMC*

En este apéndice se listan los segmentos más "gramaticales" del *Corpus del Español Mexicano Contemporáneo*. El criterio aplicado para reunirlos fue calcular la fuerza de asociación que cada forma del corpus exhibe con respecto al resto de los segmentos que allí ocurren, esto es, se midió la cantidad de glutinosidad que cada forma exhibe tanto a la derecha como a la izquierda. Cabe aclarar que no se tomó en cuenta la ocurrencia de signos de puntuación: es decir, sólo se utilizó la glutinosidad general y no la escrita.

El índice de ordenamiento utilizado para ordenar los segmentos de la tabla D.1 es sencillamente la suma de las glutinosidades estimadas para cada lado de cada segmento. La penúltima columna exhibe la diferencia absoluta de glutinosidades. Mientras mayor la diferencia, mayor el carácter que el segmento gramatical tiene de ser un clítico. Los valores negativos implican una tendencia a la encliticidad, los positivos a la procliticidad.

Nótese cómo la progresión de segmentos va de los altamente gramaticales (aquellos con los valores mayores) a los segmentos menos gramaticales (pero que obtuvieron valores de

glutinosidad todavía importantes en comparación con el resto de las formas del corpus). Como era de esperarse no es clara la frontera entre lo definitivamente gramatical y aquello que, aunque muy usado, es de carácter gramatical cuestionable. Es más, mientras que en las primeras páginas de la tabla predominan formas de significados mínimos, la mayoría de las que aparecen al final son analizables en segmentos menores. De todas maneras, es sorprendente que particularmente éstas últimas parezcan ser parte de una gramática superior a la morfosintáctica: una estructura de la cultura en México.

Tabla D.1: Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
1.	y	60303	8638	1955	-6684	10590
2.	de	114346	7002	2260	-4743	9262
3.	en	50123	6963	238.3	-6725	7201
4.	del	18621	3805	1079	-2726	4884
5.	a	45525	4145	439	-3706	4584
6.	con	18747	3532	402.8	-3129	3935
7.	que	67243	2444	1135	-1309	3579
8.	se	33623	743.7	2649	1905	3393
9.	para	14655	2639	689.9	-1949	3328
10.	el	51708	1330	1977	647.5	3307
11.	por	19835	2998	213.9	-2784	3212
12.	al	11179	1996	1177	-819.2	3173
13.	la	73110	700.6	2315	1615	3016
14.	un	19765	522.9	2045	1522	2568
15.	los	31231	311.9	2020	1708	2331
16.	una	16473	457.9	1857	1399	2315
17.	su	12520	206.7	1934	1728	2141
18.	o	8264	1583	341	-1242	1924
19.	las	20882	258.3	1324	1065	1582
20.	como	11088	1011	363.8	-646.7	1374
21.	más	9778	501.9	869.6	367.7	1371
22.	sus	5267	80.28	1153	1073	1233
23.	me	10410	75.27	1149	1074	1224
24.	muy	4915	296.2	692.2	396	988.4
25.	es	18601	600.2	316	-284.2	916.3
26.	le	8502	56.63	846.9	790.3	903.6
27.	lo	13705	151.4	682.2	530.8	833.7
28.	nos	3399	60.45	718.9	658.5	779.4
29.	no	31676	350.6	421.9	71.28	772.5
30.	son	4447	290.8	404.2	113.4	695
31.	sin	3179	288.9	347	58.14	635.9
32.	sobre	2740	573.2	20.12	-553.1	593.3
33.	tan	1591	162	430.9	268.9	592.9
34.	te	3389	34.57	535.4	500.9	570
35.	entre	2528	534.1	33.57	-500.6	567.7
36.	dos	3216	49.92	497.4	447.5	547.3
37.	esta	2925	43.08	480.5	437.4	523.6
38.	está	4108	96.51	415.7	319.2	512.2
39.	ser	2830	23.34	450.4	427	473.7
40.	les	2021	26.67	436.3	409.6	463

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
41.	mi	4840	90.97	355.1	264.2	446.1
42.	cuando	4765	240.6	188.6	-52.02	429.2
43.	qué	4523	14.82	402.9	388.1	417.7
44.	e	1325	156.5	258	101.5	414.5
45.	esa	1666	37.36	370.1	332.7	407.4
46.	puede	2480	53.11	350.3	297.1	403.4
47.	están	1561	54.81	338.8	284	393.6
48.	también	3501	259.4	134	-125.5	393.4
49.	hasta	3018	316.5	55.95	-260.6	372.5
50.	porque	5087	259.8	100.9	-158.9	360.6
51.	había	2023	26.25	328.5	302.3	354.8
52.	ya	9791	202.8	150.6	-52.11	353.4
53.	ese	1977	34.07	312.9	278.9	347
54.	pero	8336	230	107.7	-122.3	337.6
55.	gran	1182	5.579	317.8	312.2	323.4
56.	ha	4105	44.1	277.4	233.3	321.5
57.	fue	2827	138.3	171.3	32.96	309.6
58.	mucho	1754	245.8	62.16	-183.6	307.9
59.	así	4161	248.4	47.97	-200.4	296.4
60.	tu	1278	41.72	250.8	209.1	292.5
61.	bien	2603	98.52	190.8	92.28	289.3
62.	yo	7044	157.3	129.2	-28.07	286.5
63.	han	1947	27.36	257.9	230.6	285.3
64.	ni	2392	179.7	103.4	-76.21	283.1
65.	si	6122	154.1	127.1	-26.98	281.3
66.	debe	1318	70.18	183.2	113	253.4
67.	hacia	911	241.5	10.96	-230.5	252.5
68.	aquí	4024	180.7	69.89	-110.9	250.6
69.	otros	1499	9.734	240	230.3	249.8
70.	99	16449	108.6	138.2	29.61	246.9
71.	pueden	944	36.05	204.1	168.1	240.2
72.	sí	9079	183.5	51.22	-132.3	234.7
73.	tiene	3122	73.98	155	81.03	229
74.	tres	1598	32.81	195.2	162.4	228
75.	hay	4013	34.02	193.2	159.2	227.3
76.	grandes	845	31.76	192.1	160.3	223.9
77.	estas	765	12.47	211.1	198.7	223.6
78.	nada	2730	184.3	37.75	-146.5	222
79.	era	2650	84.56	130.6	46.01	215.1
80.	sido	1050	0.6681	213.4	212.8	214.1

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
81.	mis	963	26.41	181.3	154.8	207.7
82.	estos	1005	9.587	192.1	182.5	201.6
83.	otro	1979	30.02	168.9	138.9	198.9
84.	desde	1904	174.7	23.17	-151.6	197.9
85.	esos	741	6.908	186.9	180	193.8
86.	fueron	861	64.53	128.1	63.58	192.6
87.	nuestra	768	10.8	180.3	169.5	191.1
88.	estaba	1368	37.62	151.6	114	189.2
89.	otras	1071	12.41	171.4	159	183.8
90.	uno	3075	127.8	53.38	-74.42	181.2
91.	mal	579	84.73	96.23	11.51	181
92.	donde	1968	49	131.1	82.07	180.1
93.	bastante	510	70	106.3	36.27	176.3
94.	ahí	2060	88.07	84.46	-3.601	172.5
95.	muchos	941	26.47	144	117.5	170.4
96.	pues	6077	91.09	77.68	-13.4	168.8
97.	entonces	2974	88.77	79.99	-8.782	168.8
98.	unas	643	42.17	126.2	84.03	168.4
99.	mejor	1110	41.03	127.2	86.21	168.3
100.	hacer	1707	37.57	128.6	91.07	166.2
101.	mayor	1209	32.81	132.1	99.29	164.9
102.	nacional	445	127.7	34.96	-92.75	162.7
103.	algunos	822	36.93	124.9	87.95	161.8
104.	estar	665	30.25	131.3	101.1	161.6
105.	cuya	267	8.143	153.2	145	161.3
106.	siempre	1655	59.58	101.4	41.85	161
107.	unos	1166	55.06	104.8	49.72	159.8
108.	durante	1017	148.8	10.91	-137.9	159.7
109.	pa	788	113.7	45.98	-67.68	159.6
110.	este	7157	39.49	119.9	80.44	159.4
111.	buen	436	8.521	147.8	139.3	156.3
112.	misma	887	2.154	152.2	150.1	154.4
113.	será	747	76.75	76.67	-0.08411	153.4
114.	habían	443	13.06	140.1	127.1	153.2
115.	allí	973	67.41	85.08	17.66	152.5
116.	usted	1482	65.33	83.85	18.52	149.2
117.	eran	561	51.24	97.56	46.32	148.8
118.	quien	998	12.98	135.3	122.3	148.3
119.	otra	2082	50.81	93.55	42.74	144.4
120.	haber	604	12.98	131.3	118.3	144.3

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
121.	mismo	1801	4.259	139.7	135.4	143.9
122.	estoy	693	20.11	123.8	103.7	143.9
123.	algo	1109	105.9	36.04	-69.82	141.9
124.	diferentes	412	19.48	122.1	102.7	141.6
125.	hace	2003	20.8	120.8	99.96	141.6
126.	nueva	452	4.336	135.8	131.5	140.2
127.	estaban	402	28.05	111.9	83.85	140
128.	tienen	1275	25.73	112.8	87.09	138.6
129.	cuatro	833	23.27	114.7	91.43	138
130.	sólo	1716	33.87	102.6	68.78	136.5
131.	cada	1878	23.91	111.5	87.64	135.4
132.	varios	438	23.53	110.8	87.29	134.3
133.	conmigo	251	110.3	21.11	-89.23	131.5
134.	hacen	733	9.439	119.9	110.5	129.4
135.	casi	1133	61.05	66.82	5.772	127.9
136.	social	503	104.7	21.41	-83.33	126.2
137.	solo	405	18.03	108.1	90.05	126.1
138.	nuestros	445	7.278	118.3	111	125.5
139.	esas	658	11.68	113.1	101.4	124.8
140.	dentro	882	122.7	0.2712	-122.4	123
141.	buena	519	15.81	106.4	90.62	122.2
142.	usté	728	63.37	58.67	-4.708	122
143.	ahora	1923	58.99	62.94	3.956	121.9
144.	cómo	1694	56.49	64.22	7.733	120.7
145.	cualquier	655	8.051	111.4	103.4	119.5
146.	podría	367	25.99	91.58	65.59	117.6
147.	todavía	780	65.6	51.25	-14.35	116.8
148.	puedo	439	7.498	109.1	101.6	116.6
149.	hemos	523	22.71	93.67	70.97	116.4
150.	podemos	380	22.23	93.93	71.71	116.2
151.	allá	1229	59.95	56.1	-3.848	116
152.	nomás	813	51.59	63.78	12.19	115.4
153.	él	2633	19.16	96.19	77.03	115.3
154.	ella	2111	24.67	90.49	65.82	115.2
155.	mexicana	265	89.39	24.96	-64.43	114.3
156.	nuevo	544	6.652	107.1	100.4	113.7
157.	nunca	1046	47.65	65.63	17.98	113.3
158.	todo	4119	63.32	49.66	-13.65	113
159.	vida	1632	1.448	110.9	109.5	112.4
160.	nuestro	802	8.262	104	95.72	112.2

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
161.	poder	514	8.284	102.4	94.11	110.7
162.	completamente	191	44.16	63.86	19.7	108
163.	pueda	320	9.863	97.47	87.61	107.3
164.	menos	1413	13.88	93.02	79.13	106.9
165.	dar	755	11.33	95.41	84.08	106.7
166.	nuestras	283	4.929	99.05	94.12	104
167.	tenía	1100	30.32	73.48	43.16	103.8
168.	demasiado	214	40.65	63.06	22.41	103.7
169.	ante	621	92.08	11.6	-80.48	103.7
170.	tengo	1126	16.88	86.4	69.52	103.3
171.	nosotros	1126	25.52	77.58	52.06	103.1
172.	ustedes	513	28.79	74.19	45.4	103
173.	tener	793	30.79	71.53	40.75	102.3
174.	mucha	437	29.24	72.84	43.6	102.1
175.	forma	1495	4.904	96.87	91.96	101.8
176.	ellos	1459	13.68	87.85	74.16	101.5
177.	algunas	561	29.29	71.97	42.68	101.3
178.	primera	811	0.8291	100.4	99.58	101.2
179.	algún	449	18.55	82.37	63.82	100.9
180.	estamos	475	21.33	78.8	57.47	100.1
181.	ningún	357	26.9	72.38	45.48	99.28
182.	nuevos	241	21.89	76.22	54.33	98.11
183.	contra	675	83.54	13.96	-69.58	97.5
184.	primer	788	1.671	95.12	93.45	96.79
185.	he	1200	15.58	80.53	64.95	96.11
186.	hizo	737	6.174	88.76	82.59	94.93
187.	alguna	565	29.93	64.98	35.04	94.91
188.	tenemos	631	27.75	66.86	39.11	94.61
189.	tus	339	23.52	68.46	44.94	91.98
190.	mayores	237	53.8	37.52	-16.28	91.32
191.	eso	3219	26.41	64.85	38.43	91.26
192.	antes	1288	71.76	18.96	-52.81	90.72
193.	queda	405	11.93	78.21	66.28	90.14
194.	demás	483	0.2108	89.86	89.65	90.07
195.	pudo	235	12.24	77.79	65.54	90.03
196.	esto	1542	41.21	48.76	7.553	89.97
197.	haciendo	392	24.41	65.14	40.74	89.55
198.	estuvo	378	34.2	55.31	21.11	89.5
199.	propia	304	12.65	76.58	63.94	89.23
200.	luego	1957	12.26	76.93	64.67	89.19

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
201.	libre	280	70.62	18.13	-52.49	88.74
202.	bajo	627	69.95	18.18	-51.77	88.12
203.	haya	381	10.04	77.38	67.34	87.41
204.	internacional	159	75.71	11.62	-64.09	87.33
205.	siendo	355	23.64	63.66	40.02	87.3
206.	quieren	238	20.31	66.78	46.47	87.09
207.	sean	286	11.19	75.2	64.01	86.38
208.	muchas	851	42.73	42.79	0.05892	85.52
209.	mundo	949	0.8425	84.08	83.24	84.93
210.	popular	136	58.86	26	-32.86	84.85
211.	sea	1608	13.96	70.39	56.42	84.35
212.	ps	1211	14.66	68.71	54.05	83.37
213.	orita	533	39.23	44	4.765	83.22
214.	pos	1377	26.99	56.01	29.03	83
215.	deben	467	14.08	68.42	54.34	82.51
216.	ayer	434	56	26.4	-29.6	82.41
217.	mexicano	391	53.23	28.44	-24.79	81.66
218.	dónde	347	7.656	73.89	66.23	81.54
219.	mejores	234	10.91	70.12	59.2	81.03
220.	segunda	306	1.002	79.49	78.49	80.49
221.	ir	824	51.68	28.15	-23.53	79.82
222.	dinero	615	46.38	33.41	-12.97	79.8
223.	apenas	284	22.1	57.42	35.32	79.51
224.	podía	343	8.484	70.63	62.14	79.11
225.	quedó	309	10.02	69.03	59.02	79.05
226.	hoy	743	30.66	48.02	17.37	78.68
227.	quiere	731	15.89	61.98	46.08	77.87
228.	tanto	1418	29.63	47.87	18.25	77.5
229.	poco	1418	11.78	65.68	53.9	77.46
230.	somos	285	15.88	61.54	45.67	77.42
231.	soy	744	9.821	67.3	57.48	77.12
232.	varias	321	21.37	54.97	33.6	76.35
233.	hacía	323	10.83	65.17	54.34	76
234.	posible	580	17.87	57.95	40.08	75.82
235.	después	2051	59.05	16.66	-42.4	75.71
236.	años	2165	17.47	57.12	39.65	74.59
237.	mexicanos	339	46.47	28.02	-18.45	74.49
238.	segundo	444	3.958	70.49	66.53	74.45
239.	serán	209	28.99	45.06	16.07	74.05
240.	quiero	629	8.247	65.78	57.54	74.03

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
241.	hombre	1381	4.871	68.96	64.09	73.83
242.	esté	275	10.38	63.13	52.75	73.51
243.	tiempo	1878	19.17	54.02	34.86	73.19
244.	tú	1297	24.62	48.33	23.71	72.95
245.	quería	345	11.45	61.28	49.83	72.73
246.	sería	385	19.24	52.95	33.72	72.19
247.	principal	258	31.66	40.28	8.623	71.94
248.	dicha	176	5.151	66.65	61.5	71.8
249.	hubiera	362	9.735	61.73	52	71.46
250.	ellas	512	4.438	65.93	61.49	70.37
251.	va	1907	31.2	38.82	7.622	70.01
252.	da	821	7.651	62.27	54.62	69.92
253.	estás	199	15.19	54.51	39.32	69.69
254.	cosa	1135	4.527	64.88	60.36	69.41
255.	natural	261	58.29	10.95	-47.34	69.24
256.	tampoco	310	14.73	53.7	38.97	68.42
257.	día	1752	5.988	62.29	56.31	68.28
258.	aunque	944	8.379	59.64	51.26	68.02
259.	cual	867	0.6152	67.4	66.78	68.01
260.	pura	231	16.19	51.79	35.6	67.97
261.	diversas	190	4.082	63.55	59.47	67.63
262.	humana	168	54.95	12.61	-42.34	67.56
263.	aquella	340	14.5	52.31	37.81	66.81
264.	dicho	530	10.93	55.58	44.65	66.51
265.	cuyo	206	2.384	64.07	61.68	66.45
266.	aquel	398	6.824	59.51	52.69	66.34
267.	pasado	509	19.41	46.77	27.36	66.18
268.	trabajo	1297	14.13	52.02	37.9	66.15
269.	nuevas	229	12.31	53.83	41.51	66.14
270.	fuera	736	41.13	24.93	-16.2	66.06
271.	días	1079	10.77	55.15	44.38	65.92
272.	veces	1334	2.542	63.24	60.7	65.78
273.	viene	555	32.44	33.33	0.8944	65.77
274.	seguir	274	20.18	45.18	25	65.36
275.	total	483	55.2	9.974	-45.23	65.18
276.	política	476	26.89	38.16	11.27	65.05
277.	ninguna	395	20.24	44.71	24.47	64.95
278.	deberá	214	14.49	50.21	35.72	64.7
279.	tenido	300	2.457	62.1	59.65	64.56
280.	último	398	3.188	61.08	57.89	64.26

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
281.	menor	302	15.82	48.38	32.56	64.21
282.	cosas	1144	18.11	45.83	27.72	63.94
283.	puro	186	11.45	52.06	40.61	63.52
284.	tal	1169	19.22	44.05	24.83	63.28
285.	quienes	370	3.988	59.28	55.29	63.27
286.	cinco	750	23.25	39.94	16.69	63.19
287.	perfectamente	136	40.31	22.88	-17.43	63.19
288.	necesario	480	8.203	54.93	46.73	63.14
289.	ciertas	142	8.519	54.56	46.04	63.08
290.	todos	2627	40.6	22.42	-18.18	63.03
291.	sola	301	18.31	43.96	25.65	62.27
292.	feliz	162	31.41	30.79	-0.6157	62.19
293.	existen	237	11.79	50.39	38.6	62.18
294.	diversos	229	6.739	55.16	48.43	61.9
295.	hubo	374	19.49	42.29	22.8	61.78
296.	hecho	923	12.58	49.18	36.6	61.77
297.	común	262	49.19	12.39	-36.8	61.58
298.	totalmente	184	35.16	26.42	-8.741	61.58
299.	visto	344	8.332	53.08	44.75	61.41
300.	nadie	501	10.97	50.44	39.47	61.41
301.	superior	267	49.75	11.64	-38.11	61.4
302.	única	201	2.576	58.81	56.24	61.39
303.	general	850	51.19	9.927	-41.26	61.11
304.	dando	181	13.95	47.14	33.19	61.09
305.	dan	421	4.75	56.08	51.33	60.83
306.	militar	87	43.57	16.97	-26.6	60.55
307.	ojos	617	4.088	56.36	52.28	60.45
308.	actualmente	216	20.07	40.3	20.23	60.37
309.	anda	217	25.36	34.96	9.608	60.32
310.	acá	547	21.34	38.97	17.64	60.31
311.	únicamente	216	28.97	31.13	2.155	60.1
312.	tenían	267	21.12	38.11	16.98	59.23
313.	ái	490	9.637	49.36	39.73	59
314.	valor	412	10.84	47.96	37.12	58.8
315.	quizá	245	8.372	50.35	41.97	58.72
316.	cierta	184	9.625	49.04	39.42	58.67
317.	cultural	149	47.73	10.63	-37.1	58.37
318.	siquiera	244	1.485	56.34	54.86	57.83
319.	hago	202	11.22	45.86	34.64	57.08
320.	especial	311	43.44	13.23	-30.21	56.67

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
321.	profesional	113	41.93	14.72	-27.2	56.65
322.	político	187	38.57	17.98	-20.59	56.55
323.	completa	141	40.3	16.01	-24.29	56.31
324.	humano	177	38.5	17.79	-20.71	56.28
325.	mujer	1046	4.927	51.24	46.32	56.17
326.	dejan	132	6.582	49.51	42.92	56.09
327.	aún	489	18.17	37.79	19.63	55.96
328.	primeros	270	0.9714	54.87	53.9	55.85
329.	pasar	373	30.31	25.35	-4.962	55.66
330.	alto	321	8.056	47.57	39.51	55.62
331.	solamente	460	19.55	35.9	16.35	55.45
332.	datos	293	13.71	41.37	27.66	55.09
333.	verde	156	24.37	30.55	6.175	54.92
334.	pus	526	5.4	49.48	44.08	54.88
335.	productos	309	10.83	44.03	33.21	54.86
336.	baja	209	16.6	38.19	21.59	54.79
337.	alta	205	6.472	48.26	41.79	54.73
338.	real	181	43	11.66	-31.33	54.66
339.	contigo	147	41.08	13.55	-27.53	54.63
340.	sigue	346	14.83	39.78	24.95	54.6
341.	verdadera	139	3.697	50.69	47	54.39
342.	estado	609	7.285	47.07	39.79	54.35
343.	claramente	99	40.32	13.75	-26.57	54.07
344.	hicieron	285	5.955	48.1	42.14	54.05
345.	mil	825	16.19	37.72	21.53	53.91
346.	anterior	460	30.92	22.98	-7.941	53.9
347.	nacionales	149	46.69	7.201	-39.49	53.89
348.	vienen	352	30.73	23.12	-7.613	53.85
349.	país	877	1.414	52.2	50.79	53.61
350.	actual	285	31.82	21.77	-10.06	53.59
351.	junto	321	52.94	0.5489	-52.39	53.49
352.	ahorita	224	19.29	34.04	14.75	53.32
353.	grande	613	33.23	19.77	-13.45	53
354.	éste	607	12.41	40.59	28.18	53
355.	agua	1360	7.412	45.57	38.16	52.98
356.	realmente	242	18.74	34.17	15.42	52.91
357.	bueno	2328	31.43	21.33	-10.1	52.75
358.	pequeños	182	23.28	29.44	6.154	52.72
359.	carácter	316	2.427	50.05	47.62	52.47
360.	saben	168	11.1	41.34	30.24	52.44

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
361.	abierto	99	39.56	12.87	-26.7	52.43
362.	industrial	167	38.2	14.06	-24.14	52.25
363.	podrá	184	9.272	42.97	33.7	52.24
364.	escrito	134	29.46	22.69	-6.774	52.15
365.	partes	373	14.32	37.83	23.51	52.15
366.	jamás	198	3.772	48.25	44.47	52.02
367.	tenga	270	6.727	45.24	38.52	51.97
368.	público	421	11.36	40.58	29.22	51.94
369.	hacerlo	214	20.97	30.81	9.84	51.77
370.	siguientes	250	0.7409	50.97	50.23	51.71
371.	grupos	280	18.34	33.35	15	51.69
372.	normal	182	39.51	12.03	-27.48	51.53
373.	andan	104	33.04	18.49	-14.55	51.53
374.	países	474	7.709	43.79	36.08	51.5
375.	sociedad	272	1.779	49.67	47.9	51.45
376.	comer	352	4.152	47.23	43.08	51.38
377.	tierra	674	4.021	47.05	43.03	51.07
378.	última	260	2.437	48.51	46.07	50.95
379.	información	340	4.004	46.32	42.31	50.32
380.	pesos	947	8.078	42.16	34.08	50.23
381.	gente	955	7.328	42.76	35.44	50.09
382.	inicial	90	38.84	11.16	-27.68	49.99
383.	pequeñas	153	19.75	30.19	10.44	49.95
384.	seis	475	21.46	28.49	7.029	49.95
385.	horas	617	14.32	35.62	21.3	49.93
386.	haga	231	9.224	40.62	31.39	49.84
387.	pueblo	717	1.716	48.12	46.4	49.84
388.	quieres	313	6.212	43.6	37.39	49.82
389.	frío	191	23.57	26.19	2.623	49.76
390.	personal	289	32.65	17.05	-15.6	49.7
391.	cincuenta	305	40.45	9.092	-31.36	49.55
392.	mala	160	13.77	35.72	21.95	49.49
393.	joven	359	12.08	37.39	25.31	49.47
394.	dio	432	2.969	46.48	43.51	49.45
395.	media	468	10.63	38.79	28.15	49.42
396.	viendo	167	29.12	20.19	-8.931	49.31
397.	directamente	145	30.19	18.98	-11.21	49.17
398.	mañana	734	12.29	36.86	24.57	49.15
399.	loco	91	34.27	14.62	-19.65	48.89
400.	ciertos	140	10.07	38.8	28.72	48.87

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
401.	necesita	259	7.639	41.2	33.56	48.83
402.	corazón	341	2.104	46.58	44.48	48.69
403.	ora	464	14.85	33.82	18.97	48.67
404.	quedan	162	4.583	43.99	39.41	48.58
405.	saber	431	22.9	25.66	2.756	48.56
406.	tercera	141	1.317	47.22	45.9	48.53
407.	constante	161	38.13	10.28	-27.85	48.41
408.	vale	123	14.16	34.23	20.08	48.39
409.	muerto	187	31.27	17.03	-14.24	48.3
410.	económico	307	28.55	19.72	-8.836	48.27
411.	vivo	128	30.92	17.34	-13.59	48.26
412.	namás	165	22.86	25.22	2.359	48.08
413.	fácil	239	11.65	36.3	24.65	47.95
414.	moderna	122	30.05	17.76	-12.29	47.81
415.	inferior	135	36.58	11.09	-25.49	47.68
416.	sentido	565	16.48	31.1	14.63	47.58
417.	partido	202	10.79	36.68	25.89	47.47
418.	problemas	532	23.56	23.89	0.3317	47.45
419.	debemos	180	5.614	41.8	36.19	47.42
420.	hacerse	183	15.26	32.14	16.88	47.39
421.	hice	157	8.772	38.56	29.79	47.33
422.	cultura	292	3.666	43.39	39.72	47.05
423.	negra	99	33.47	13.58	-19.9	47.05
424.	buscando	106	30.13	16.88	-13.25	47.02
425.	mismos	323	1.027	45.79	44.77	46.82
426.	ambos	255	3.999	42.79	38.79	46.79
427.	conocido	135	27.32	19.47	-7.85	46.79
428.	casa	1512	2.72	44.06	41.34	46.78
429.	condiciones	469	10.83	35.79	24.96	46.62
430.	relativamente	81	4.582	41.97	37.39	46.55
431.	tienes	376	13.43	33.11	19.67	46.54
432.	mujeres	445	9.371	37.11	27.74	46.48
433.	diez	554	28.85	17.57	-11.27	46.42
434.	económica	240	37.64	8.664	-28.97	46.3
435.	igual	505	35.33	10.95	-24.38	46.28
436.	-	2190	17.71	28.51	10.8	46.22
437.	ahi	79	11.67	34.5	22.84	46.17
438.	hacían	125	7.853	38.3	30.45	46.16
439.	distintos	131	8.877	37.23	28.35	46.11
440.	pequeño	215	12.89	33.14	20.25	46.04

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
441.	toda	1191	22.02	23.98	1.962	46
442.	cuidado	199	30.33	15.59	-14.74	45.92
443.	puedes	156	8.091	37.76	29.67	45.85
444.	tarde	662	5.718	40.12	34.4	45.83
445.	eres	372	13.77	31.98	18.21	45.76
446.	pobre	232	15.57	30.13	14.56	45.69
447.	salió	245	22.04	23.64	1.595	45.68
448.	fundamentales	81	42.56	3.091	-39.47	45.66
449.	presente	319	13.63	31.92	18.29	45.55
450.	salir	432	22.06	23.45	1.38	45.51
451.	próximo	215	8.147	37.34	29.19	45.48
452.	hijos	522	9.971	35.49	25.51	45.46
453.	elementos	399	17.94	27.47	9.532	45.42
454.	personas	707	13.67	31.72	18.05	45.39
455.	madre	470	3.653	41.68	38.02	45.33
456.	sentir	110	19.93	25.34	5.402	45.27
457.	blanca	144	34.65	10.6	-24.05	45.25
458.	baile	215	10.21	35.03	24.82	45.23
459.	comercial	111	33.57	11.55	-22.03	45.12
460.	hablar	435	29.78	15.34	-14.45	45.12
461.	estudios	318	9.997	35.08	25.08	45.08
462.	central	122	37.51	7.568	-29.94	45.07
463.	trabajando	244	24.79	20.22	-4.57	45.02
464.	exactamente	212	19.02	25.99	6.965	45.01
465.	vez	1944	1.162	43.78	42.62	44.94
466.	dado	410	11.9	32.95	21.06	44.85
467.	ideas	251	10.24	34.44	24.2	44.68
468.	oscuro	70	24.81	19.74	-5.066	44.55
469.	abajo	238	16.41	28.03	11.63	44.44
470.	empleado	71	23.87	20.5	-3.371	44.37
471.	viejo	209	18.99	25.23	6.231	44.22
472.	mercado	264	1.139	42.92	41.78	44.06
473.	propio	302	3.786	40.12	36.33	43.91
474.	pasó	260	12.01	31.83	19.82	43.85
475.	seguido	121	20.78	23.05	2.271	43.84
476.	primeras	136	0.5167	43.31	42.79	43.83
477.	van	877	23.86	19.88	-3.982	43.73
478.	andaba	124	21.75	21.91	0.1654	43.66
479.	hacemos	130	12.27	31.29	19.02	43.57
480.	comerciales	98	36.37	7.179	-29.19	43.55

Tabla D.1 (continuación):
Formas gramaticales del *CEMC*

	forma	fr.	glutinosidad de un lado	glutinosidad del otro	diferencia	índice de ordenamiento
481.	cuesta	80	12.85	30.67	17.82	43.53
482.	papá	572	1.646	41.88	40.23	43.52
483.	comprar	171	4.121	39.28	35.15	43.4
484.	buenas	239	16.13	27.24	11.11	43.38
485.	precisamente	355	17.69	25.67	7.977	43.36
486.	sale	360	19.13	24.1	4.971	43.24
487.	resulta	221	14.16	29.05	14.9	43.21
488.	tomar	341	12.35	30.79	18.45	43.14
489.	has	271	8.433	34.68	26.25	43.12
490.	ganado	200	7.129	35.95	28.82	43.08
491.	pasaron	75	23.96	18.83	-5.135	42.79
492.	artística	72	29.01	13.77	-15.24	42.78
493.	actividad	269	7.31	35.45	28.14	42.76
494.	ti	300	0.7875	41.85	41.06	42.63
495.	quisiera	186	14.68	27.93	13.24	42.61
496.	onde	140	11.91	30.66	18.75	42.57
497.	sistema	637	3.13	39.44	36.31	42.57
498.	siente	152	1.706	40.85	39.14	42.55
499.	vivir	327	12.18	30.36	18.19	42.54
500.	distintas	137	10.69	31.84	21.16	42.53

Bibliografía

[1] Abbagnano, Nicola, tr. Alfredo N. Galletti. *Diccionario de filosofía.* Fondo de Cultura Económica. México, 2ª ed., 1991 [1961].

[2] Aho, A.V. y J.D. Ullman. *The Theory of Parsing. Translation, and Compiling.* Prentice-Hall. Nueva York, 1972.

[3] Alarcos Llorach, Emilio. *Gramática de la Lengua Española.* Espasa Calpe, Madrid, 1ª edic., 1999 [1994].

[4] Allen, James. *Natural Language Understanding.* Benjamin/Cummings. Redwood, California, 2ª ed., 1995.

[5] Altmann-Fitter. Iterative fitting of probability distributions. Lüdenscheid: RAM, 1997.

[6] Altmann, Gabriel. *Statistik für Linguisten*, vol. 55 de *Quantitative Linguistics.* Wissenschaftlicher Verlag, Trier, 1995.

[7] Altmann, Gabriel. Reseña: Michael Oakes, *Statistics for Corpus Linguistics. Journal of Quantitative Linguistics*, 6(3):269–270, 1999.

[8] Anderson. Stephen R. *A-Morphous Morphology*, vol. 62 de *Cambridge Studies in Linguistics.* Cambridge University Press, Cambridge, 1994.

[9] Anis, Jacques. "¿Una grafemática autónoma?", págs. 271–285, *Hacia una teoría de la lengua escrita.* [31], 1986.

[10] Antworth, Evan L. *PC-KIMMO: A Two-Level Processor for Morphological Analysis*, vol. 16 de *Occasional Publications in Academic Computing.* Summer Institute of Linguistics, Dallas, 1990.

[11] Arnush, Craig. *Teach Yourself Borland C++ 5.* Sams Publishing, Indianapolis, 3 ed., 1996.

[12] Baayen, Rolf Harald. *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation.* Akademisch proefschrift, Vrije Universitait te Amsterdam, Amsterdam, 1989.

[13] Barnbrook, Geoff. *Language and Computers. A Practical Introduction to the Computer Analysis of Language.* Edinburgh University Press, Edinburgh, 1998 [1996].

[14] Bátori, István S., Winfred Lenders, y Wolfgang Putschke, eds. *Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications.* Walter de Gruyter, Berlín/Nueva York, 1989.

[15] Battestini, Simon. "Escrituras africanas (inventario y problemática", págs. 195–205. *Hacia una teoría de la lengua escrita.* [31], 1986.

[16] Bello, Andrés. *Gramática de la lengua castellana.* Editorial Sopena Argentina, Buenos Aires, 5ª ed., 1953.

[17] Bergenholtz, Henning y Joachim Mugdan. *Einführung in die Morphologie.* Kohlhammer, Stuttgart, 1979.

[18] Borland International. C++ Development Suite for Windows 95, NT, 3.1 and DOS. version 5.01, 1996.

[19] Borland International. *C++ Language Reference*, vol. 5 de *Borland C++ version 5.* Scotts Valley, California, 1996.

[20] Borland International. *C++ Programmer's Guide*, vol. 2 de *Borland C++ version 5.* Scotts Valley, California, 1996.

[21] Borland International. *C++ User's Guide*, vol. 1 de *Borland C++ version 5.* Scotts Valley, California, 1996.

[22] Bright, William, ed. *The International Encyclopedia of Linguistics.* Oxford University Press, Oxford, 1992.

[23] Buenrostro Díaz, Elsa Cristina. Corpus de la lengua chuj. Archivo plano de 86KB. 2002.

[24] Bunge, Mario. *Scientific Research I. The Search for System*, vol. 3/I de *Studies in the Foundations, Methodology and Philosophy of Science.* Springer-Verlag, Berlin/Heidelberg, 1967.

[25] Bunge, Mario. *Scientific Research II. The Search for Truth*, vol. 3/II de *Studies in the Foundations, Methodology and Philosophy of Science.* Springer-Verlag, Berlin/Heidelberg, 1967.

[26] Bunge, Mario. *Philosophy of Science I: From Problem to Theory.* Transaction Publishers, New Brunswick, 1998 [1967].

[27] Bunge, Mario. *Philosophy of Science II: From Explanation to Justification.* Transaction Publishers, New Brunswick, 1998 [1967].

[28] Bußmann, Hadumod. *Lexikon der Sprachwissenschaft.* Kröner, Stuttgart, 2ª ed., 1990.

[29] Bybee, Joan. *Morphology. A Study of the Relation between Meaning and Form.* vol. 9 de *Typological Studies in Language.* John Benjamins, Amsterdam, 1985.

[30] Catach, Nina. "La escritura como plurisistema, o teoría de L prima", págs. 310–331. *Hacia una teoría de la lengua escrita.* [31]. 1986.

[31] Catach, Nina, ed., tr. Lía Varela y Patricia Willson. *Hacia una teoría de la lengua escrita.* Gedisa, Barcelona, 1996.

[32] Charniak, Eugene. *Statistical Language Learning.* The MIT Press, Cambridge (Mass.). 1993.

[33] Chomsky, Noam. "On Certain Formal Properties of Grammars". *Information and Control,* 2:137–167, 1959.

[34] Chomsky, Noam. "Formal Properties of Grammars", págs. 323–418, *Handbook of Mathematical Psychology.* Vol. II de Luce, R.D. *et al.* [92], 1963.

[35] Chomsky, Noam. *Cartesian Linguistics. A Chapter in the History of Rationalist Thought.* University Press of America, Lanham, 1966.

[36] Chomsky, Noam y George A. Miller. "Finite State Languages". *Information and Control,* 1:91–112, 1958.

[37] Chomsky, Noam y George A. Miller. "Introduction to the Formal Analysis of Natural Languages", págs. 269–322, *Handbook of Mathematical Psychology.* Vol. II de Luce, R.D. *et al.* [92], 1963.

[38] Chomsky, Noam y George A. Miller. "Finitary Models of Language Users", págs. 419–492, *Handbook of Mathematical Psychology.* Vol. II de Luce, R.D. *et al.* [92], 1963.

[39] Chomsky, Noam, tr. Carlos-Peregrín Otero. *Estructuras sintácticas.* Siglo XXI, México, 11ª ed., 1994 [1974 [1957]].

[40] Church, Kenneth W. y Robert L. Mercer. "Introduction to the Special Issue on Computational Linguistics Using Large Corpora". *Computational Linguistics,* 19:1–24, 1993.

[41] Company Company, Concepción y Alfonso Medina Urrea. "Sintaxis motivada pragmáticamente. Futuros analíticos y futuros sintéticos". *Revista de Filología Española,* (LXXIX):65–100, 1999.

[42] Company, Concepción. "Los futuros en el español medieval. Sus orígenes y evolución". *Nueva Revista de Filología Española,* 34:48–107, 1985-86.

[43] Cooper, W. D., tr. Jairo Panesco Tascon. *Instrumentación electrónica y mediciones.* Prentice Hall, Bogotá, 1982 [1970].

[44] Corominas. Joan y José A. Pascual. *Diccionario crítico etimológico castellano e hispánico. vol. I-VI.* vol. 7 de *Diccionarios.* Gredos. Madrid. 1991 [1980].

[45] Coulmas. Florian. "Superación de la diglosia: acercamiento del japonés escrito y hablado en el siglo XIX", págs. 242–256. *Hacia una teoría de la lengua escrita.* [31]. 1986.

[46] Cromm. Oliver. *Affixerkennung in deutschen Wortformen. Eine Untersuchung zum nicht-lexikalischen Segmentierungsverfahren von N. D. Andreev.* Abschluß des Ergänzungsstudiums Linguistische Datenverarbeitung. Francfort del Meno. 1996.

[47] Fernández García, Javier. *Acerca de la teoría de la información y algunas de sus aplicaciones*, vol. 23 de *Comunicación Interna.* Departamento de Matemáticas, Facultad de Ciencias. UNAM, México, 1978.

[48] Flenner. Gudrun. "Ein quantitatives Morphsegmentierungssystem für spanische Wortformen", págs. 31–62, *Computation Linguae II.* Vol. 83 de Klenk, Ursula [77]. 1994.

[49] Frakes, William B. "Stemming Algorithms", págs. 131–160. *Information Retrieval. Data Structures and Algorithms.* Prentice Hall. New Jersey. 1992.

[50] García Hidalgo, María Isabel. "La formalización del analizador gramatical del DEM", págs. 85–155, *Investigaciones lingüísticas en lexicografía.* Vol. 89 de *Jornadas* [89], 1a. edición. 1979.

[51] Gazdar. Gerald, Edwan Klein, Goeffrey Pullum. y Ivan A. Sag. *Generalized Phrase Structure Grammar.* Harvard University Press. Cambridge (Mass.). 1985.

[52] Gazdar. Gerald y Chris Mellish. *Natural Language Processing in Prolog.* Addison-Wesley. Wokingham. Gran Bretaña. 1989.

[53] Gelbukh, Alexander, ed. *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002.* vol. 2276 de *Lecture Notes in Computer Science.* Springer, México, febrero 2002.

[54] Gevarter. William B., übers. Susanne Bader. *Intelligente Maschinen. Einführung in die künstliche Intelligenz und Robotik.* VCH. Weinheim. 1985.

[55] Gili Gaya. Samuel. *Curso superior de sintaxis española.* Vox/Bibliograf. Barcelona. 15ª ed.. 1994.

[56] Glück. Helmut, ed. *Metzler Lexikon Sprache.* Verlag J.B. Metzler. Stuttgart. 2ª ed.. 2000.

[57] Gonzalvo Mainar, Gonzalo. *Diccionario de metodología estadística.* Ediciones Morata. Madrid. 1978.

[58] Greenberg, Joseph H. *Essays in Linguistics.* The University of Chicago Press, Chicago. 1967 [1957].

[59] Haarmann, Harald. *Universalgeschichte der Schrift.* Campus/Parkland, Francfort del Meno, 1998 [1991].

[60] Hafer, Margaret A. y Stephen F. Weiss. "Word Segmentation by Letter Successor Varieties". *Information Storage and Retrieval.* 10:371–385, 1974.

[61] Hallebeek, Jos, tr. Pieter de Haan. *A Formal Approach to Spanish Syntax,* vol. 7 de *Language and Computers: Studies in Practical Linguistics.* Editions Radopi, Amsterdam/Atlanta, 1992.

[62] Halpern, Aaron L. "Clitics", págs. 101–122, *The Handbook of Morphology.* [130]. 1998.

[63] Ham Chande, Roberto y Luis Fernando Lara. "Del 1 al 100 en lexicografía", págs. 41–83. *Investigaciones lingüísticas en lexicografía.* Vol. 89 de *Jornadas* [89], 1a. edición. 1979.

[64] Harris, Zellig S. "Morpheme Alternants in Linguistic Analysis". *Language,* 18:169–180, 1942.

[65] Harris, Zellig S. "From Phoneme to Morpheme". *Language,* 31(2):190–222, 1955.

[66] Harris, Zellig S. *A Theory fo Language and Information. A Mathematical Approach.* Clarendon Press, Oxford, 1991.

[67] Hockett, Charles F. Reseña: Shannon y Weaver, *The Mathematical Theory of Communication. Language,* 29:69–93, 1953.

[68] Hockett, Charles F. "Linguistic Elements and their Relations". *Language,* 37:29–53, 1961.

[69] Hopcroft, John E. y John D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading (Mass.), 1979.

[70] Janßen, Axel. "Segmentierung französischer Wortformen in Morphe ohne Verwendung eines Lexikons", págs. 74–95, *Computation Linguae I.* Vol. 73 de Klenk, Ursula [75], 1992.

[71] Jäppinen, Harri. "Finite State Computational Morphology", págs. 96–109, *Computation Linguae I.* Vol. 73 de Klenk, Ursula [75]. 1992.

[72] Jiménez, Héctor y Guillermo Morales. "SEPE: A POS Tagger for Spanish", págs. 250–259, *Computational Linguistics and Intelligent Text Processing.* Vol. 2276 de *Lecture Notes in Computer Science* [53], febrero 2002.

[73] Juilland, Alphonse y E. Chang Rodríguez. *A Frequency Dictionary of Spanish Words*. Mouton, La Haya, 1965.

[74] Kageura, Kyo. "Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences". *Journal of Quantitative Linguistics*, 6:149–160, 1999.

[75] Klenk, Ursula, ed. *Computation Linguae I*, vol. 73 de *ZDL-Beiheft*. Franz Steiner, Stuttgart, 1992.

[76] Klenk, Ursula. "Verfahren morphologischer Segmentierung und die Wortstruktur des Spanischen", págs. 110–124, *Computation Linguae I*. Vol. 73 de *ZDL-Beiheft* [75], 1992.

[77] Klenk, Ursula, ed. *Computation Linguae II*, vol. 83 de *ZDL-Beiheft*. Franz Steiner, Stuttgart, 1994.

[78] Klenk, Ursula. "Automatische morphologische Analyse arabischer Verbformen", págs. 84–101, *Computation Linguae II*. Vol. 83 de *ZDL-Beiheft* [77], 1994.

[79] Klenk, Ursula y Hagen Langer. "Morphological Segmentation Without a Lexicon". *Literary and Linguistic Computing*, 4(4):247–253, 1989.

[80] Kock, Josse de y Walter Bossaert. *Introducción a la lingüística automática en las lenguas románicas*, vol. 202 de *Estudios y Ensayos*. Gredos, Madrid, 1974.

[81] Kock, Josse de y Walter Bossaert. "De la definición de estructuras lingüísticas con la ayuda de un ordenador. El morfema", págs. 181–227, *Introducción a la lingüística automática en las lenguas románicas*. Vol. 202 de *Estudios y Ensayos* [80], 1974.

[82] Kock, Josse de y Walter Bossaert. *The Morpheme. An Experiment in Quantitative and Computational Linguistics*. Van Gorcum, Amsterdam, Madrid, 1978.

[83] Köhler, Reinhard. "Diversification of Coding Methods in Grammar", págs. 47–55, *Diversification Processes in Language: Grammar*. [121], 1991.

[84] Koskenniemi, Kimmo. "Computational Morphology", págs. 291–293, *International Encyclopedia of Linguistics*. Bright, William [22], 1992.

[85] Lara, Luis Fernando. *Dimensiones de la lexicografía. A propósito del Diccionario del español de México*, vol. 116 de *Jornadas*. El Colegio de México, A.C., México, 1a. edición, 1990.

[86] Lara, Luis Fernando. "Caracterización metódica del *corpus* del DICCIONARIO DEL ESPAÑOL DE MÉXICO", págs. 85–106, *Dimensiones de la lexicografía. A propósito del Diccionario del español de México*. Vol. 116 de *Jornadas* [85], 1a. edición, 1990.

[87] Lara, Luis Fernando. "La cuantificación en el DICCIONARIO DEL ESPAÑOL DE MÉXICO", págs. 51–84. *Dimensiones de la lexicografía. A propósito del Diccionario del español de México.* Vol. 116 de *Jornadas* [85]. 1a. edición. 1990.

[88] Lara, Luis Fernando y Roberto Ham Chande. "Base estadística del Diccionario del Español de México", págs. 5–39, *Investigaciones lingüísticas en lexicografía.* Vol. 89 de *Jornadas* [89]. 1a. edición, 1974.

[89] Lara, Luis Fernando, Roberto Ham Chande, y Ma. Isabel García Hidalgo. *Investigaciones lingüísticas en lexicografía*, vol. 89 de *Jornadas*. El Colegio de México, A.C., México, 1a. edición, 1979.

[90] Lara, Luis Fernando, dir. *Diccionario del español usual de México.* El Colegio de México A.C., México, 1a. edic., 1996.

[91] Lázaro Carreter, Fernando. *Diccionario de términos filológicos*, vol. 6 de *Manuales*. Gredos, Madrid. 3ª ed., 1990.

[92] Luce, R.D., R. Bush, y E. Galanter, eds. *Handbook of Mathematical Psychology, Vol. I-III.* Wiley, Nueva York, 1963.

[93] Manning, Christopher D. y Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge (Mass.), 1999.

[94] Mason, Oliver. *Programming for Corpus Linguistics: How to do Text Analysis with Java.* Edinburgh University Press, Edinburgh, 2000.

[95] Matthews, P. H. *Morphology.* Cambridge University Press, Cambridge, second edition, 1991.

[96] Medina Urrea, Alfonso. "Un experimento cuantitativo de determinación de fronteras morfológicas del español de México". IV Encuentro Internacional de Lingüística en el Noroeste, Hermosillo, Sonora, noviembre 1996.

[97] Medina Urrea, Alfonso. "Automatic Discovery of Affixes by Means of a Corpus: A Catalog of Spanish Affixes". *Journal of Quantitative Linguistics*, 7(2):97–114, 2000.

[98] Medina Urrea, Alfonso. "Propiedades lingüístico-cuantitativas de cadenas de caracteres (segmentos, palabras, vocablos) en c$\frac{3}{4}$rpora de lenguajes naturales: *afijalidad* y *cliticidad* en el español de México". VI Encuentro Internacional de Lingüística en el Noroeste, Hermosillo, Sonora, noviembre-diciembre 2000.

[99] Medina Urrea, Alfonso. "Características cuantitativas de la morfología flexiva del chuj". VII Encuentro Internacional de Lingüística en el Noroeste, Hermosillo, Sonora, noviembre 2002.

[100] Meillet, A. *Linguistique historique et linguistique générale.* La Société de Linguistique de Paris, Paris, 1958.

[101] Meillet, A. "L'évolution des formes grammaticales", págs. 130–148. *Linguistique historique et linguistique générale.* [100], 1958 [1912].

[102] Meya, Montserrat. "Morphologische Analyse des Spanischen", págs. 134–156. *Informationslinguistische Texterschließung.* Vol. 4 de Schwarz, Christoph y Thurmair [123], 1986.

[103] Meyer-Eppler, W. *Grundlagen und Anwendungen der Informationstheorie.* vol. 1 de *Kommunikation un Kybernetik in Einzeldarstellungen.* Springer Verlag, Heidelberg, 2ª ed., 1969.

[104] Moliner, María. *Diccionario de uso del español,* vol. 5 de *Biblioteca Románica Hispánica. Diccionarios.* Gredos, Madrid, 1992.

[105] Moreno de Alba, José. *Morfología derivativa nominal en el español de México.* UNAM, 1986.

[106] Moreno de Alba, José. *La preficación en el español mexicano.* UNAM, 1996.

[107] Mounin, Georges, tr. Felisa Marcos. *Historia de la lingüística desde los orígenes al siglo* XX. vol. 16 de *Manuales.* Gredos, Madrid, 1989 [1967].

[108] Naumann, Sven y Hagen Langer. *Parsing: Eine Einführung in die maschinelle Analyse natürlicher Sprache.* Teubner, Stuttgart, 1994.

[109] Nida, Eugene A. "The Identification of Morphemes". *Language,* 24, 1948.

[110] Nida, Eugene A. *Morphology. The Descriptive Analysis of Words.* The University of Michigan Press, Ann Arbor, 1967 [1949].

[111] Oakes, Michael P. *Statistics for Corpus Linguistics.* Edinburgh University Press, Edinburgh, 1998.

[112] Pereira, Fernando C. N. y Stuart M. Shieber. *Prolog and Natural-Language Analysis,* vol. 10 de *CSLI Lecture Notes.* Center for the Study of Language and Information, Stanford, 1987.

[113] Piotrowski, R. G., K. B. Bektaev, y A. A. Piotrowskaja, übersetzt von A. Falk. *Mathematische Linguistik,* vol. 27 de *Quantitative Linguistics.* Brockmeyer, Bochum, 1985.

[114] Piotrowski, R. G., M. Lesohin, y K. Lukjanenkov. *Introduction of Elements of Mathematics to Linguistics,* vol. 44 de *Quantitative Linguistics.* Brockmeyer, Bochum, 1990.

[115] Porter, M.F. "An Algorithm for Suffix Stripping". *Program,* 14(3):130–137, 1980.

[116] Rainer, Franz. *Spanische Wortbildungslehre.* Niemeyer, Tübingen, 1993.

[117] Rey-Debove, Josette. "En busca de la distinción oral-escrito", págs. 97–115. *Hacia una teoría de la lengua escrita.* [31], 1986.

[118] Reynolds, Leighton D. y Nigel G. Wilson, tr. Manuel Sánchez Mariana. *Copistas y filólogos*, vol. 20 de *Manuales*. Gredos, Madrid, 1986 [1974].

[119] Rini, Joel. *Motives for Linguistic Change in the Formation of the Spanish Object Pronouns*. Juan de la Cuesta, Newark, Delaware, 1992.

[120] Ritchie, Graeme D., Graham J. Russell, Alan W. Black, y Stephen G. Pulman. *Computational Morphology. Practical Mechanisms for the English Lexicon*. The MIT Press, Cambridge (Mass.), 1992.

[121] Rothe, Ursula, ed. *Diversification Processes in Language: Grammar*. Margit Rottman Medienverlag, Hagen, 1991.

[122] Sapir, Edward, tr. Margit y Antonio Alatorre. *El lenguaje*, vol. 96 de *Breviarios*. Fondo de Cultura Económica, México, 1992 [1921].

[123] Schwarz, Christoph y Gregor Thurmair, eds. *Informationslinguistische Texterschließung*, vol. 4 de *Linguistische Datenverarbeitung*. Georg Olms Verlag, Zürich, 1986.

[124] Searle, John R. "Minds, Brains, and Programs". *The Behavioral and Brain Sciences*, 3:417–457, 1980.

[125] Shannon, Claude E. y Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.

[126] Shieber, Stuart M. *An Introduction to Unification-Based Approaches to Grammar*, vol. 4 de *CSLI Lecture Notes*. Center for the Study of Language and Information, Stanford, 1986.

[127] Shieber, Stuart M. *Constraint-Based Grammar Formalisms. Parsing and Type Inference for Natural and Computer Languages*. The MIT Press, Cambridge (Mass.)/London, 1992.

[128] Slocum, Jonathan. "Machine Translation: A Survey of Active Systems", págs. 629–645. István S. Bátori, Winfred Lenders and Wolfgang Putschke, *Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications*. Walter de Gruyter, Berlín/Nueva York, 1989.

[129] Spencer, Andrew. *Morphological Theory. An Introduction to Word Structure in Generative Grammar*. Basil Blackwell, Cambridge, 1991.

[130] Spencer, Andrew y Arnold M. Zwicky. *The Handbook of Morphology*. Blackwell, Oxford, 1998.

[131] Sproat, Richard William. *Morphology and Computation*. The MIT Press, Cambridge (Mass.), 1992.

[132] Thurmair, Gregor. "Ein Morphologisches Prozessorsegment zur Erzeugung von Grundformen mithilfe von Lernverfahren". págs. 8–31. *Informationslinguistische Texterschließung.* Vol. 4 de Schwarz, Christoph y Thurmair [123]. 1986.

[133] Varrón, Marco Terencio, tr. Manuel-Antonio Marcos Casquero. *De lingua Latina.* vol. 6 de *Textos y Documentos Clásicos del Pensamiento y de las Ciencias.* Anthropos, Madrid, 1990 [*ca.* 40 a.C.].

[134] Weaver, Warren. "Recent Contributions to the Mathematical Theory of Communication", págs. 3–28, *The Mathematical Theory of Communication.* [125]. 1964.

[135] Wells, Rulon S. "Automatic Alternation". *Language*, 25:99–116, 1949.

[136] Woods, Anthony, Paul Fletcher, y Arthur Hughes. *Statistics in Language Studies.* Cambridge University Press. 1986.