

Fundamentos de tipología lingüística

PROYECTO EDITORIAL CLAVES DE LA LINGÜÍSTICA

Director:
Juan Carlos Moreno Cabrera



Queda prohibida, salvo excepción prevista en la ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con autorización de los titulares de la propiedad intelectual. La infracción de los

derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (arts. 270 y sigs. Código Penal). El Centro Español de Derechos Reprográficos (www.cedro.org) vela por el respeto de los citados derechos.

Fundamentos de tipología lingüística

Iván Igartua



EDITORIAL
SÍNTESIS

Consulte nuestra página web: www.sintesis.com
En ella encontrará el catálogo completo y comentado

Los mapas incluidos en este libro se han obtenido del *World atlas of language structures online* (WALS Online), editado por M. S. Dryer y M. Haspelmath, haciendo uso de la herramienta para la adaptación de símbolos y del aspecto visual de los propios mapas.

Reservados todos los derechos. Está prohibido, bajo las sanciones penales y el resarcimiento civil previstos en las leyes, reproducir, registrar o transmitir esta publicación, íntegra o parcialmente, por cualquier sistema de recuperación y por cualquier medio, sea mecánico, electrónico, magnético, electroóptico, por fotocopia o por cualquier otro, sin la autorización previa por escrito de Editorial Síntesis, S. A.

© Iván Igartua

© EDITORIAL SÍNTESIS, S. A.
Vallehermoso, 34. 28015 Madrid
Teléfono: 91 593 20 98
www.sintesis.com

ISBN: 978-84-1357-263-5
Depósito Legal: M. 10.424-2023

Impreso en España - Printed in Spain

Índice

Prólogo	11
Abreviaturas	13
1. Introducción: diversidad lingüística y tipología	17
1.1. La diversidad lingüística: diversidad de lenguas, diversidad filogenética y diversidad estructural	17
1.2. Factores que influyen en la diversidad lingüística	21
1.3. Diversidad entre lenguas	23
1.3.1. <i>Lenguas naturales y lenguas artificiales</i>	23
1.3.2. <i>Lenguas orales y lenguas de signos</i>	24
1.3.3. <i>Lenguas de transmisión ininterrumpida y lenguas de contacto</i>	25
1.3.4. <i>Complejidad estructural</i>	28
1.4. La diversidad puesta (equivocadamente) en cuestión	31
1.5. ¿Qué es, por tanto, y a qué se dedica la tipología?	34
1.6. Vertientes de la tipología	38
1.6.1. <i>Tipología sincrónica y tipología diacrónica</i>	38
1.6.2. <i>Tipología parcial y tipología holística</i>	38
1.6.3. <i>Macrotipología y microtipología</i>	39
1.7. Tipos de tipología	39
1.7.1. <i>Tipología areal</i>	40
1.7.2. <i>Tipología sociolingüística</i>	40
1.7.3. <i>Tipología intragenética</i>	41
1.8. Tipología y otras disciplinas lingüísticas	41
1.9. Breve historia de los estudios tipológicos	42
2. Principios y bases metodológicas de la tipología	51
2.1. Los límites de la tipología (y, por extensión, de la lingüística)	51
2.2. Muestras de lenguas	54

2.2.1. Muestras de conveniencia	54
2.2.2. Muestras de probabilidad y de variedad	55
2.3. Base empírica de la investigación tipológica y comparabilidad ...	62
2.4. Conceptos comparativos y categorías descriptivas	63
2.5. ¿Qué hay de la teoría?	66
2.6. Recursos y herramientas para los estudios tipológicos	69
2.7. Conclusiones	72
3. Universales, generalizaciones tipológicas y jerarquías	73
3.1. El concepto de universal y clases de universales	73
3.2. Los cuadros tetracóricos	80
3.3. Tendencias generales y rarezas tipológicas	84
3.4. Jerarquías tipológicas	86
3.4.1. La jerarquía referencial, nominal o de animacidad	87
3.4.2. La jerarquía de accesibilidad o de relaciones gramati- cales	89
3.4.3. La jerarquía de concordancia	90
3.4.4. La jerarquía de incorporación	91
3.5. Conclusiones: lo que puede haber detrás de los universales lingüísticos	93
4. Tipología fonológica	97
4.1. Los inventarios fonológicos	98
4.2. Tipología segmental: el vocalismo	99
4.2.1. Sistemas vocálicos	99
4.2.2. Monoptongos y diptongos	103
4.2.3. Vocales orales y nasales	103
4.2.4. Armonía vocálica	106
4.3. Tipología segmental: el consonantismo	107
4.4. Ratio consonante/vocal	116
4.5. Tipología suprasegmental o prosódica	117
4.5.1. Cantidad	117
4.5.2. Estructura silábica	118
4.5.3. Acento	122
4.6. Fonología de las lenguas de signos	125
4.7. Perspectivas ecológicas sobre la tipología fonológica	127
4.8. Conclusiones	129
5. Tipología morfológica	131
5.1. Clases de morfemas	132
5.2. Flexión y derivación	133

Índice

5.3.	Parámetros de la clasificación morfológica	136
5.3.1.	<i>Fusión morfológica</i>	137
5.3.2.	<i>Densidad semántica</i>	139
5.3.3.	<i>Alomorfia</i>	142
5.3.4.	<i>Grado de síntesis</i>	144
5.4.	Diversidad de las relaciones entre significante y significado	145
5.5.	Complejidad morfológica	150
5.6.	La afijación	154
5.6.1.	<i>Clases de afijos</i>	154
5.6.2.	<i>Tendencias tipológicas en la afijación</i>	155
5.6.3.	<i>Afijos y clíticos</i>	158
5.7.	La morfología de los morfemas	159
5.8.	Morfología de las lenguas de signos	161
5.9.	Conclusiones	164
6.	<i>Tipología de las categorías gramaticales</i>	165
6.1.	El género	166
6.1.1.	<i>Organización de la categoría</i>	166
6.1.2.	<i>Número de géneros</i>	168
6.1.3.	<i>Sistemas de asignación del género</i>	169
6.1.4.	<i>El género en los pronombres</i>	172
6.1.5.	<i>Clases nominales y clasificadores</i>	173
6.2.	El número	176
6.2.1.	<i>Organización de la categoría</i>	176
6.2.2.	<i>Valores del número</i>	178
6.2.3.	<i>Medios de expresión del número</i>	181
6.2.4.	<i>El número verbal</i>	184
6.2.5.	<i>Relaciones entre número y género</i>	185
6.3.	El caso	187
6.3.1.	<i>Sistemas de caso</i>	187
6.3.2.	<i>Medios de expresión del caso</i>	193
6.4.	El tiempo	195
6.4.1.	<i>Valores del tiempo</i>	196
6.4.2.	<i>Tiempo métrico</i>	196
6.4.3.	<i>Tiempo nominal</i>	197
6.5.	El aspecto	199
6.5.1.	<i>Aspecto verbal y modo de acción (Aktionsart)</i>	200
6.5.2.	<i>Medios de expresión del aspecto</i>	201
6.5.3.	<i>Aspectualidad nominal</i>	202
6.6.	El modo y las modalidades	203
6.7.	La evidencialidad: ¿cómo se cuenta la historia?	205

6.7.1. Valores de la evidencialidad	205
6.7.2. Medios de expresión de la evidencialidad	207
6.8. La persona	209
6.8.1. Organización de la categoría	209
6.8.2. Medios de expresión de la persona	212
6.9. Categorías gramaticales en las lenguas de signos	215
6.10. Conclusiones	217
7. Tipología sintáctica	219
7.1. El orden básico de palabras	220
7.1.1. Cuestiones preliminares	220
7.1.2. Órdenes posibles y su presencia en las lenguas del mundo	221
7.1.3. Razones de las preferencias universales	225
7.1.4. El orden de otros constituyentes	227
7.2. Relaciones gramaticales y alineamiento	229
7.2.1. Alineamiento del sujeto	230
7.2.2. Alineamiento del objeto	237
7.2.3. Marcaciones diferenciales	240
7.2.4. Marcación de núcleo y de dependiente	248
7.3. La concordancia	249
7.3.1. Tipos de concordancia	251
7.3.2. Rasgos que participan en la concordancia	252
7.3.3. La jerarquía de concordancia	254
7.4. Oraciones complejas	256
7.4.1. Coordinación	257
7.4.2. Subordinación	259
7.4.3. Cosubordinación	267
7.5. Sintaxis de las lenguas de signos	271
7.6. Conclusiones	273
8. Tipología léxica y semántica	275
8.1. Las clases de palabras	276
8.2. Distinciones entre clases de palabras	277
8.3. Clases funcionales de palabras	281
8.3.1. Pronombres	282
8.3.2. Adposiciones	285
8.3.3. Numerales	285
8.3.4. Artículos	288
8.3.5. Auxiliares	289
8.3.6. Conjunciones	291

Índice

8.3.7. <i>Interjecciones</i>	293
8.4. Taxonomía y partonomía	293
8.4.1. <i>Partes del cuerpo</i>	294
8.4.2. <i>Términos de parentesco</i>	295
8.4.3. <i>Designaciones de los colores</i>	296
8.5. La búsqueda de primitivos semánticos	298
8.6. Los mapas semánticos	300
8.7. Clases léxico-semánticas de palabras en las lenguas de signos ..	302
8.8. Conclusiones	304
9. Tipología holística	305
9.1. Bases de la tipología holística	305
9.2. Propuestas de clasificación holística en la tipología moderna ...	308
9.2.1. <i>Los tipos de Skalička</i>	308
9.2.2. <i>El principio estructural de Lehmann</i>	309
9.2.3. <i>La tipología rítmica de Donegan y Stampe</i>	310
9.2.4. <i>La tipología prosódica de Gil</i>	311
9.2.5. <i>La tipología sistémica y la complejidad silábica de</i> <i>Fenk-Oczlon y Fenk</i>	311
9.2.6. <i>El prototipo aislante-monocategorial-asociativo de</i> <i>Gil</i>	313
9.3. Valoración de las hipótesis holísticas	314
9.4. Tendencias del cambio lingüístico desde una perspectiva holís- tica	315
9.5. Conclusiones	320
10. Tipología diacrónica	323
10.1. La evolución comparada de las lenguas	324
10.2. Usos inadecuados de la tipología en propuestas diacrónicas	325
10.3. Estabilidad tipológica e inferencias históricas	327
10.4. Tendencias generales de la evolución lingüística	328
10.4.1. <i>La evolución fonológica</i>	329
10.4.2. <i>Tendencias y mecanismos de la evolución morfológica</i> ..	333
10.4.3. <i>Tendencias y mecanismos de la evolución sintáctica</i> ...	339
10.4.4. <i>Orden sintáctico y orden morfológico</i>	343
10.4.5. <i>El cambio léxico y semántico</i>	344
10.4.6. <i>La gramaticalización</i>	346
10.5. Contacto lingüístico	349
10.5.1. <i>El efecto de las áreas lingüísticas</i>	351
10.5.2. <i>La metatipia</i>	352
10.6. Tipología sociolingüística	353

10.7. Tipología y reconstrucción	354
10.8. Tipología del cambio y cambio de la tipología	358
10.8.1. Cambios morfológicos tipológicamente relevantes	358
10.8.2. Contacto de lenguas y cambio tipológico	364
10.8.3. Simplificación y complejización	365
10.9. La evolución de las lenguas de signos	366
10.10. Conclusiones	369
11. La explicación en tipología	371
11.1. Las preferencias universales	372
11.2. La hipótesis innatista	372
11.3. La explicación funcionalista (o funcional-adaptativa)	373
11.3.1. Economía y frecuencia	375
11.3.2. Iconicidad	377
11.3.3. Imitación y analogía	378
11.3.4. Eficiencia del procesamiento	379
11.3.5. Conflictos entre motivaciones	380
11.3.6. Críticas a la explicación funcional	382
11.4. La explicación histórica	384
11.5. Conclusiones: la función tras la historia y el uso	388
Bibliografía esencial	391

2

Principios y bases metodológicas de la tipología

La tipología lingüística es, en primer lugar, una aproximación teórica y metodológica especial a los fenómenos lingüísticos. En su base está la investigación empírica de esos fenómenos, a partir de la cual pueden construirse, mediante inducción, generalizaciones acerca de las preferencias de estructuración que presentan las lenguas, preferencias entendidas como tendencias compartidas (en un grado u otro). Pero su foco esencial se encuentra en el estudio de la distribución de las diversas unidades o estructuras lingüísticas, en las posibilidades de variación que presenta la expresión de determinadas relaciones gramaticales o semánticas así como su materialización fonológica. Para ello, la tipología ha desarrollado en los últimos años diversos métodos de análisis, adaptados a los objetivos de cada tipo de estudio. Esa metodología y los principios que animan la investigación tipológica de las lenguas forman el núcleo de este capítulo.

2.1. Los límites de la tipología (y, por extensión, de la lingüística)

Es un hecho evidente que no hay estudio tipológico posible sin la comparación de lenguas. La tipología solo existe en la medida en que se conocen y se han descrito diversas lenguas, cuyos elementos o estructuras formales pueden cotejarse. Tras la comparación de fenómenos lingüísticos, una de las consecuencias suele ser la identificación de elementos hasta cierto punto similares, patrones más o menos comunes que se asocian a tendencias generales en la configuración fonológica o gramatical de las lenguas, lo que en tipología tradicional se conoce como universales

(sean bien absolutos, bien estadísticos, cf. para ello el capítulo 3). Pero enseguida surge una primera cuestión de relevancia para la investigación en tipología: ¿cómo se llega a este tipo de generalizaciones, a esos universales? Es decir: ¿en qué medida se justifica la formulación de esas preferencias lingüísticas? En palabras de Song (2001: 17), “intuitively speaking, the best way to discover language universals is perhaps to examine all languages of the world”. Desde luego, el estudio de todas las lenguas del mundo sería no solo el mejor camino, sino, en rigor, la única vía posible para extraer conclusiones válidas acerca de las tendencias generales en la estructuración de las lenguas.

Pero hay, lamentablemente, un obstáculo que no es en absoluto despreciable: de las más de 7 000 lenguas (o bien no más de 4 000, según otras estimaciones, cf. Dixon 2010-2012, I: xiii) que se hablan hoy día en el mundo, solo una pequeña porción ha sido descrita hasta la fecha y no siempre de manera conveniente, esto es, de un modo que permita una comparación en profundidad de la mayor parte de sus características gramaticales (de ahí que algunos investigadores consideren que el conjunto de lenguas convenientemente descritas no supera el 10%, cf. Evans y Levinson 2009: 432, Song 2018: 32). Al ritmo actual de documentación y análisis de lenguas, se necesitarán todavía unos cuantos decenios –si no siglos– para llegar a describir la mayor parte de las lenguas hoy existentes (o las que queden en el momento de su documentación y descripción). Por otro lado, según un cálculo de Bakker (2011: 101), basado a su vez en estimaciones de Nettle (1999a: 100 ss.) que proyectan la tasa de densidad lingüística actual a épocas del pasado, en los últimos 40 000 años puede que se hayan llegado a hablar unas 240 000 lenguas (incluidas las actuales) por distintas comunidades de *homo sapiens* (considerando que en cada momento habría alrededor de 6 000 lenguas y que cada una de ellas duraría unos 1 000 años). Las lenguas a las que se tiene acceso hoy día solo representan un 3% de ese conjunto (y las que están descritas, un 0,95%). En opinión de algunos lingüistas (Evans y Levinson 2009), el panorama sería aún peor. La formulación, a partir de estos datos, de universales del lenguaje o, lo que es lo mismo, de propiedades o reglas lingüísticas que no admiten excepción, resulta una tarea ya no solo altamente complicada y arriesgada, sino carente de todo sentido estadístico: el conjunto de lenguas que conocemos relativamente bien es tan ínfimo con respecto al número de lenguas que existen o han existido que las generalizaciones de carácter absoluto no encuentran soporte. Ni siquiera la aplicación del llamado principio de uniformidad (*uniformitarianism or uniformity principle*, cf. Lass 1997: 24-32) o del *actualismo* (como lo denominaba Luis Michelena en la estela de geólogos y biólogos) permite salvar la situación: aunque el principio se asume en general (siempre y cuando, se podría añadir, la distancia temporal no sea excesiva), hay razones evolutivas para pensar que algunas propiedades o dinámicas de sistemas lingüísticos hablados hace varias decenas de miles de años no tienen por qué coincidir enteramente con las que caracterizan a las lenguas actuales (Comrie 1993, Heine y Kuteva 2007).

Podría considerarse, siendo pesimistas, que estas limitaciones minan severamente las aspiraciones ya no solo de la tipología como disciplina, sino por extensión de la propia lingüística en tanto ciencia. Desde luego, todo depende de la trascendencia que se quiera otorgar a las conclusiones que se extraen de cada investigación y, por tanto, del propio enfoque que se adopte en el análisis, lo que, en cualquier caso, no impide la formulación siempre legítima de hipótesis provisionales sobre la base del material estudiado (Song 2018: 32). Pero también es cierto que hay otras ciencias que se encuentran en situaciones hasta cierto punto equiparables, cuando no sensiblemente peores, y que no por ello dejan de cultivarse y desarrollarse. Un ejemplo particularmente ilustrativo es la biología: en la actualidad hay catalogadas más de 1 500 000 especies animales (algunos elevan la cifra hasta 1 800 000), pero se calcula que tal vez existan nueve millones (según estimaciones más bien conservadoras, dado que algunos biólogos piensan que la realidad puede estar más cerca de los 23 o incluso 30 millones de especies, *vid.* Mayr 2002: 293). Este sería el panorama actual, al que habría que añadir todo el inmenso conjunto de organismos animales –como los que originó la llamada explosión cámbrica de hace algo más de 530 millones de años– que han existido en el pasado y de los que, en algunos casos, no quedan más que recuerdos fósiles (algunos, en realidad muchos, linajes evolutivos simplemente quedaron truncados, cf. Gould 2006). La extinción masiva de la era cretácico-paleógena hace aproximadamente 65 millones de años, asociada popularmente a la desaparición de los dinosaurios, ocasionó la aniquilación de cerca del 76 % de los géneros biológicos que habitaban la Tierra en aquella era. Con anterioridad hubo al menos otros siete episodios de extinción masiva (con resultados cercanos en algún caso al 95 % de desaparición de las especies, como a finales de la era pérmica, hace unos 245 millones de años, cf. Mayr 2002: 223). En resumen, si la biología puede sostenerse como ciencia pese a estas indudables limitaciones, no menos debería decirse de la lingüística.

La realidad es, en cualquier caso, que no cabe diseñar estudios tipológicos que analicen todas las lenguas actuales. Está en primer lugar la imposibilidad de acceso a toda la información requerida, dado que solo una porción de las lenguas se encuentra adecuadamente descrita, esto es, cuenta con una gramática publicada, además de un diccionario y de una colección de textos en ciertos casos. Por otra parte, incluso en el caso hipotético de que se contara con información acerca de todas las lenguas, la tarea de manejar esos datos excedería en la actualidad los límites de cualquier estudio tipológico, por restringido que sea (más aún cuando es realizado por una sola persona o por un grupo reducido de investigadores). En estas circunstancias, la tipología opera por necesidad como lo hacen las proyecciones sociológicas: a través de instrumentos como las encuestas. En el caso de la tipología, estableciendo subconjuntos (muestras) de lenguas que aspiran a ser representativas de la diversidad lingüística global, al menos de la documentada. La metodología del muestreo se expone en la sección siguiente.

2.2. Muestras de lenguas

El muestreo de lenguas (*language sampling*), basado en la definición de un subconjunto representativo de ellas, es distinto según el objetivo de la investigación tipológica y ha ido evolucionando, en general, a lo largo de los últimos años. Se suele distinguir entre muestras o listas de conveniencia, listas aleatorias, listas de diversidad o variedad y listas probabilísticas, que se exponen a continuación.

2.2.1. Muestras de conveniencia

Las primeras muestras en tipología fueron listas de conveniencia (*convenience samples*), establecidas en función de las posibilidades bibliográficas o bien de los gustos e intuiciones del tipólogo. Una de esas listas es la que utilizó Greenberg (1963) en su artículo fundacional (cf. *supra* §1.9) sobre el orden básico de los constituyentes en la oración. Contenía treinta lenguas (cf. el cuadro 2.1) de distintas regiones del mundo, entre ellas algunas europeas como el galés, el noruego, el italiano, el finés o el euskera, varias habladas en África (en su mayor parte nígero-congoleñas, como el suajili, el yoruba o el fula), algunas de Asia (desde el turco al hindi, el canarés o kannada, el malayo o el birmano), dos en representación de Australia y Oceanía (loritja o luritja, maorí) y un puñado de lenguas amerindias (quechua, zapoteco, chibcha o guaraní).

Cuadro 2.1. Lista de conveniencia de Greenberg (1963), por continentes

Europa	África	Asia	América	Oceanía
euskera	bereber	birmano	chibcha	loritja
finés	fula	burushaski	guaraní	maorí
galés	masái	hebreo	maya	
griego mod.	nubio	hindi	quechua	
italiano	songhai	japonés	zapoteco	
noruego	suajili	kannada		
serbio	yoruba	malayo		
		tailandés		
		turco		

Las regiones mayores del mundo estaban representadas en la muestra, aunque en distinta medida: con una presencia equilibrada en términos cuantitativos de lenguas europeas, asiáticas y africanas, y una representación menor en el caso de las lenguas habladas en América y, sobre todo, de las de Australia y Oceanía (solamente dos, y ninguna de Papúa Nueva Guinea). Desde luego, en una lista así no podían tener cabida todas las familias lingüísticas, habida cuenta de que son unas

250 (a las que habría que sumar las lenguas aisladas, en torno a 159, cf. Campbell 2017: 1, 15). Pese a esas limitaciones de partida, Greenberg pudo formular 45 generalizaciones tipológicas, llamadas entonces universales, que siguen teniendo valor heurístico pese a que prácticamente todas ellas presentan excepciones (del concepto de universal y de sus clases se tratará más abajo, en §3.1). No todos esos universales tenían el mismo alcance, algo de lo que era consciente Greenberg, que los formuló al menos de tres maneras:

1. Como universales absolutos (n.º 3: las lenguas con orden dominante VSO son siempre preposicionales; n.º 34: ninguna lengua tiene un número trial salvo que tenga un dual. Ninguna tiene un dual salvo que tenga un plural).
2. Como universales casi absolutos (n.º 41: si en una lengua el verbo sucede tanto al sujeto como al objeto nominal como orden dominante, la lengua en cuestión tendrá casi siempre un sistema de casos).
3. Como universales estadísticos de un alcance inferior a los anteriores (n.º 4: con una frecuencia abrumadora, netamente mayor a la que sería producto de la casualidad, las lenguas con orden básico SOV son posposicionales).

Aunque incluso los universales formulados como absolutos están sujetos a la presencia de excepciones (el orden básico VSO no impide la presencia de posposiciones, *vid.* Song 2018: 23), sigue siendo sorprendente el grado de aproximación a generalizaciones tipológicas significativas que alcanzó Greenberg partiendo de una lista de conveniencia de solo 30 lenguas. Porque, en la actualidad, una muestra que aspire a captar gran parte de la diversidad lingüística del mundo, y ser por tanto representativa, no incluye, por lo general, menos de 250 o 300 lenguas, es decir, un número próximo al menos a la cantidad de familias lingüísticas conocidas (la media de lenguas examinadas en cada entrada del *WALS* es, por ejemplo, de 417).

2.2.2. Muestras de probabilidad y de variedad

Tanto las muestras llamadas de probabilidad (*probability samples*) como las de variedad o diversidad (*variety samples*) están basadas fundamentalmente en el control del factor genético, que puede producir en las listas uno de los sesgos más graves (junto con el areal o geográfico). Desde esta perspectiva, una lista tipológica ha de ser equilibrada en relación con las familias lingüísticas del mundo: la condición principal que ha de cumplir una muestra de variedad es que ninguna de las familias debe estar sobrerrepresentada.

La diferencia entre las muestras de probabilidad y las de variedad es sutil. En las primeras, lo que se busca es la probabilidad de que una lengua sea de un determinado tipo o posea un valor específico de un rasgo dentro de un abanico de posibilidades generalmente conocido de antemano: por ejemplo, que las lenguas

tiendan a un tipo de orden básico de palabras frente a los demás. En las segundas, se trata de explorar las opciones estructurales –aún no conocidas en su integridad– para un determinado rasgo o categoría, es decir se trata de identificar el espacio de variación en la comparación interlingüística. Un ejemplo podría ser el estudio de los medios de expresión del caso gramatical en las lenguas del mundo (Dryer 2013b), para el que se recurre típicamente a los afijos y los clíticos, pero que también puede incluir variaciones internas del tema e incluso el uso del tono como medio exclusivo de expresión del caso (en lenguas nilóticas y algunas afroasiáticas, nígero-congoleñas y tibeto-birmanas; peculiaridad recogida en el Gabinete de rarezas tipológicas de la Universidad de Constanza, cf. Moravcsik 2007: 36).

Tanto en las listas de probabilidad como en las de variedad, el control sobre el sesgo genético es fundamental, aunque por motivos distintos (e incluso, podría añadirse, en grado también distinto): en las muestras probabilísticas ese control es su propia razón de ser, dado que un eventual desequilibrio en la representación de las familias lingüísticas tendría un efecto directo sobre la estimación de la probabilidad del valor de un rasgo en concreto. En las muestras de variedad el control sobre el sesgo genético es también el punto de partida, pero el foco esencial es la consideración de la diversidad tipológica, puesto que el objetivo aquí no es medir la presencia estadística de unas variantes u otras sino identificar el mayor número de variantes posible, valores de un rasgo o medios de expresión de una categoría (intentando no pasar por alto ninguna de las posibilidades que se realizan en las lenguas). En este caso, se estima que cuanto mejor esté representada la diversidad filogenética, mayores serán las opciones de captar la diversidad estructural (con independencia del hecho constatado de que dentro de las propias familias lingüísticas puede registrarse también variedad).

La línea de análisis probabilístico es la que sigue Bell (1978), en la que puede considerarse como primera contribución monográfica al muestreo en tipología. El autor parte de la existencia de 478 grupos genéticos de varios milenios de antigüedad, algunos de ellos incluidos en lo que tradicionalmente se denomina familia lingüística. Las lenguas indoeuropeas, por ejemplo, aparecen representadas con doce grupos dentro de ese conjunto de 478, por lo que una muestra completa con esa cifra debería incluir doce lenguas de esa familia. Si la muestra fuera, en cambio, de 250 lenguas, serían seis las lenguas indoeuropeas representadas en ella. En muestras menores no todos los grupos (e incluso familias) pueden incluirse, pero Bell no establece criterios de selección para decidir cuáles formarían parte de la lista y cuáles no (cf. Bakker 2011: 111).

Uno de los métodos mejor conocidos en la confección de muestras probabilísticas es el que, siguiendo a Bell, diseñó Dryer (1989) a partir de su concepción de *genus* o conjunto de lenguas genéticamente emparentadas equiparable en antigüedad a alguna de las grandes ramas indoeuropeas (lenguas germánicas o itálicas, por ejemplo). Los *genera* –un total de 322– en que clasifica la diversidad lingüística mundial corresponden aproximadamente a las lenguas (protolenguas) que pudo haber hace unos 3 000 o 3 500 años (estimación que ya hacía Bell para sus grupos

de lenguas, cf. Bell 1978: 146-149). Sus descendientes son las lenguas y dialectos que encontramos hoy día. Familias con muchas lenguas como la bantú o la oceánica (austronesia), pero de diversificación muy reciente, tienen en la muestra una presencia limitada que no resulta proporcional al número total de lenguas.

Una vez establecidos esos *genera*, el método selecciona una lengua histórica representante de cada uno de ellos, salvo para aquellos casos en los que existe disparidad dentro del propio *genus* (entonces cada opción tipológica viene representada por una lengua perteneciente al mismo *genus*). Dryer distribuye los conjuntos de lenguas por macroáreas geográficas (inicialmente cinco, posteriormente seis): Eurasia, África, América del Norte, América del Sur, Sudeste Asiático con Oceanía y Australia con Papúa Nueva Guinea. El cuadro 2.2 recoge un ejemplo ilustrativo:

Cuadro 2.2. Distribución geográfica de la correlación entre VO y la posición de las oraciones relativas

	<i>África</i>	<i>Eurasia</i>	<i>AsiaSE</i>	<i>AusNG</i>	<i>AmN</i>	<i>AmS</i>	<i>Total</i>
VO & NRel	[30]	[9]	[13]	[4]	[14]	[5]	75
VO & RelN	0	0	1	0	0	0	1

Para Dryer, una preferencia tipológica solo es robusta (o universal) cuando, pese a las excepciones —que pueden ser más o menos numerosas— esa tendencia se detecta en todas las macroáreas geográficas del mundo (como ocurre en el ejemplo del cuadro 2.2). Puede darse la circunstancia de que una de las macroáreas tienda a la configuración no preferida: se abre en esos casos una vía para la exploración tipológica en términos de análisis diacrónico y areal, es decir, una investigación de las pautas evolutivas que conducen a la excepción en una zona determinada.

El método de Dryer establece de esta forma un control doble sobre el sesgo genético (limitando el número de lenguas de cada familia lingüística que conforman el subconjunto escogido) y sobre el sesgo geográfico o areal (identificando preferencias generales no sujetas a efectos regionales y deslindándolas precisamente de aquellos fenómenos que sí muestran una dependencia geográfica). Un punto débil del método ideado por Dryer y en general de las muestras probabilísticas es la selección “manual” de la lengua que representa cada *genus* o grupo lingüístico. Sus rasgos pueden no resultar en realidad representativos del conjunto de ese grupo y afectar por tanto a la correlación de datos. Para salvar ese escollo, otros autores (cf. Janssen, Bickel y Zúñiga 2006) han apostado por la aplicación de mecanismos de randomización, consistentes en la comparación de los resultados observados con aquellos que podrían ser producto de una distribución casual, para lo que se ha de simular mediante ordenador un alto número de distribuciones aleatorias de variables. Solo así, defienden, puede estimarse la relevancia de los datos observados, en la me-

dida en que se distancian de una distribución meramente accidental, lo que puede además conducir a conclusiones acerca de los efectos del factor genético o del areal.

Uno de los métodos de muestreo más influyentes (relacionado en este caso con la búsqueda del mayor grado de diversidad en las estructuras lingüísticas) es el que desarrollan Rijkhoff y Bakker (1998; cf. también Bakker 2011) sobre la base del índice DV (*diversity value*) de cada familia lingüística. El valor de diversidad representa el grado de complejidad en la ramificación de las familias lingüísticas, complejidad que se refleja en la cantidad de nódulos o niveles de separación de cada árbol genealógico. La fórmula que emplean para calcular ese DV incluye la ponderación de los nódulos superiores –los de mayor antigüedad dentro del árbol– con respecto a los inferiores. El resultado es un valor numérico que se proyecta después sobre la muestra que quiera emplearse en la investigación tipológica (50, 100, 200, 300 o 500 lenguas, por ejemplo). Las diferencias en el DV pueden ser muy significativas: en una de las simulaciones de Rijkhoff y Bakker, la familia urálica de lenguas presenta un índice de 11,29 y la indoeuropea un 28,53, mientras que la nígero-congoleña llega a un 55,93, la australiana al 56,42 y la indo-pacífica, macrofamilia controvertida, por lo demás, al 112,45. Ese valor se refleja en la cantidad de lenguas de cada familia lingüística que habrían de incluirse en la muestra, en función del grado de diversidad de cada una de ellas. La selección nuevamente “manual” de las lenguas representantes no está sujeta en este caso a los riesgos que presenta en las listas probabilísticas: su función es la de reflejar la variedad tipológica, no la probabilidad de un tipo u otro (lo cual no quiere decir, por otra parte, que esté libre del riesgo de pasar por alto información tipológicamente relevante).

El mínimo que se ha de cubrir en la muestra equivale siempre a un representante por cada familia (salvo que el tamaño de la muestra lo impida de entrada). Las lenguas aisladas (con un DV de 1.00) han de figurar en todo caso (aunque esto ya no es así en algunas versiones más recientes del modelo). La proyección del DV de varios grupos lingüísticos sobre muestras de diverso tamaño se recoge en el cuadro 2.3 (adaptado de Rijkhoff y Bakker 1998: 282 ss., que se basan en Voegelin y Voegelin 1977).

De este modo, la técnica de muestreo basada en el DV trata no solo de asegurar el equilibrio genético, procurando la representación de todas las familias lingüísticas y de las lenguas aisladas, sino también de simbolizar el distinto grado de diversidad que puede apreciarse en unas familias lingüísticas y otras, favoreciendo así la amplitud de análisis que se le supone a una muestra de variedad. El contraste entre la presencia de las lenguas nígero-cordofanas y la de las austronesias ilustra ese aspecto diferencial del método: pese a contar con un número mayor de lenguas, las nígero-cordofanas (un total de 1 036, según Voegelin y Voegelin 1977) están representadas por un conjunto constantemente menor que las austronesias (con 778 lenguas según la misma fuente), consecuencia de que su DV , derivado del número de nódulos del árbol genealógico, es inferior (55,93 frente a 65,57). En una muestra de 200 lenguas, las lenguas nígero-cordofanas contarían con 16 representantes y las austronesias, con 19. En una de 500, las primeras tendrían 45

puestos por 53 de las segundas. La interpretación inmediata de ese contraste es la relativa modernidad de las ramificaciones dentro de la familia nígero-cordofana (o nígero-congoleña). Las familias con un DV moderado o bajo (caso de la joisana o la esquimo-aleutiana) incrementan su representación, cuando lo hacen, de manera muy paulatina. En el caso de la esquimo-aleutiana, se vería representada solamente por una lengua tanto en una muestra de 50 o 100 lenguas como en una de 500 (hasta no llegar a la muestra de 1 000 no aparecerían hasta tres lenguas de la familia).

Cuadro 2.3. Presencia de algunas familias lingüísticas en muestras distintas (adaptado de Rijkhoff y Bakker 1998)

<i>Familias</i>	<i>Tamaño de la muestra</i>						
	<i>50</i>	<i>100</i>	<i>200</i>	<i>300</i>	<i>400</i>	<i>500</i>	<i>1 000</i>
Afroasiática	1	3	7	12	16	20	41
Ainu	1	1	1	1	1	1	1
Andina-ecuatoriana	1	5	12	19	25	32	65
Australiana	1	7	17	26	36	46	92
Austronesia	1	8	19	30	41	53	107
Burushaski	1	1	1	1	1	1	1
Esquimo-aleutiana	1	1	1	1	1	1	3
Indoeuropea	1	4	9	13	18	23	47
Indo-pacífica	1	13	33	52	71	91	184
Joisana	1	1	1	2	3	3	6
Nígero-cordofana	1	7	16	26	35	45	91
Nilo-sahariana	1	3	7	11	14	18	37
Sino-tibetana	1	3	6	10	14	17	35
Euskera	1	1	1	1	1	1	1

La principal dificultad que enfrenta esta técnica es el grado desigual de desarrollo de los estudios genéticos. Un análisis deficiente o incompleto de las relaciones de parentesco dentro de una familia lingüística (subgrupos) condiciona el grado de diversidad que se ha de plasmar posteriormente en las muestras tipológicas. Además, las muestras de variedad basadas en el DV dependen decisivamente de la calidad de las clasificaciones genéticas: algunas de ellas, como la de Ruhlen, que sirve de base a una de las simulaciones de Rijkhoff y Bakker, son sencillamente inadmisibles (según esa clasificación, que hoy día ningún lingüista serio sostiene, en el mundo solo habría 27 familias lingüísticas, cf. Campbell y Poser 2008). El cuadro 2.3 anterior está basado en otra clasificación genética, la de Voegelin y

Voegelin (1977), que, aun siendo algo más razonable, tampoco sería aceptada hoy día por la mayor parte de los lingüistas.

El modelo de Rijkhoff y Bakker no establece control sobre el sesgo regional o areal, algo que ha tratado de corregirse en propuestas posteriores, como la de Miestamo (2005). Entendiendo que la variedad estructural puede depender también de los efectos del contacto (que tienden a reducirla), Miestamo trata de conjugar las virtudes de los modelos previos (fundamentalmente el de Dryer y el basado en el DV) en su GM (*genus-macroarea sampling method*), que en esencia incorpora al muestreo una estimación de la diversidad areal, de modo que la representación de cada macroárea resulte equilibrada en función del grado de divergencia genética (calculada en *genera*) que la caracteriza. Cada macroárea recibe así un porcentaje de representación que luego se proyecta sobre el tamaño de la muestra (en esta versión, los *genera* son 521, dato derivado del WALS), como resume el cuadro 2.4 (Miestamo, Bakker y Arppe 2016: 257):

Cuadro 2.4. Distribución de lenguas según diversidad genética y areal

Macroáreas	Genera	%	Tamaño de la muestra				
			50	100	200	300	500
África	74	14,2	7	14	28	43	71
Eurasia	43	8,3	4	8	17	25	42
Sudeste Asiático y Oceanía	66	12,7	6	13	25	38	64
Australia y Nueva Guinea	140	26,9	13	27	54	81	135
América del Norte	92	17,7	9	18	35	53	89
América del Sur	106	20,3	10	20	41	61	102
Total	521	100	49	100	200	301	503

Nuevamente aquí, la mayor diversidad tanto genética como areal (número de grupos lingüísticos distintos en cada región) hace que Australia y Papúa Nueva Guinea (con un 26,9 % de la diversidad total) tengan una representación mayor de *genera* (y, por tanto, de lenguas individuales) que otras zonas del mundo: África, con un número también muy elevado de lenguas, solo tiene el 14,2% de la diversidad.

En la práctica tipológica de los últimos años lo habitual es encontrar muestras de lenguas, en general amplias (de 100 a 500 lenguas), que en realidad son listas de conveniencia (Bakker 2011: 106) y están condicionadas por las posibilidades de acceso a las fuentes de información, aunque tienden a buscar un equilibrio responsable tanto en el plano genético como en el geográfico (sin someterse, con todo, a técnicas formalizadas como aquella a la que recurre el método del valor de diversidad). Concebidas en general con función de muestras de variedad a fin de captar

la diversidad estructural, a la postre suelen ser utilizadas también en cierta medida para extraer consecuencias de alcance estadístico.

Hay finalmente un aspecto de las muestras de variedad sobre el que merece la pena reflexionar: si es cierto, como defiende Nettle (1999b: 133 ss.), que algunos rasgos peculiares e incluso extremadamente raros –por ejemplo, los órdenes básicos de palabras OVS y OSV– se documentan exclusivamente en lenguas con menos de 3 000 hablantes y que, por lo tanto, el tamaño de las comunidades lingüísticas tiene repercusión en las propiedades tipológicas de las lenguas, las muestras de variedad deberían seguramente tener una presencia sobredimensionada de lenguas con pocos hablantes (al margen de los controles previos sobre el sesgo genético). De lo contrario, puede que una parte significativa de la diversidad estructural escape a nuestra observación.

Algo similar ocurre con los enclaves geográficamente aislados (por lo común, montañosos o selváticos), en los que tienden a concentrarse propiedades tipológicas peculiares que en otras regiones no se documentan. El hecho de que esas rarezas estructurales sean prácticamente exclusivas de zonas aisladas puede llevar a pensar, por contraste, que las propiedades más extendidas en las lenguas son más bien el resultado de la difusión de rasgos, fenómeno consustancial al contacto lingüístico, y no tanto el reflejo de unos principios de estructuración compartidos e incluso inmanentes. Si la distribución de esos rasgos fonológicos y gramaticales es un hecho accidental debido a la historia de los movimientos de población, la posibilidad de alcanzar una explicación general de las estructuras lingüísticas se ve claramente perjudicada. Hoy día la presencia de clics (cf. el capítulo 4) es considerada una propiedad fonológica excepcional de las lenguas joisanas (y de algunas bantúes): pero si, frente a lo realmente ocurrido, ese grupo de comunidades joisanas hubiera proliferado por gran parte de África e incluso de Asia, el rasgo fonológico mencionado no sería tratado en la actualidad como una rareza. El carácter accidental de la evolución humana y, por tanto, de las lenguas representa un problema serio para la lingüística, entre otras disciplinas (Daniel 2007).

Desde esa perspectiva, el muestreo de lenguas, válido a efectos descriptivos, está, no obstante, limitado a la hora de facilitar una base para la explicación del hecho lingüístico, es decir, de las causas que determinan por qué las lenguas humanas son como son y se caracterizan por unos rasgos estructurales y no otros. Los filtros genético y areal que incorporan las muestras, incluso los de las más sofisticadas, no mejoran sustancialmente a este respecto la situación, sencillamente porque hay factores evolutivos generales (y contingentes) que no pueden integrar. En cualquier caso, se trata de unos límites que no resultan ajenos a otras disciplinas empíricas que han de enfrentarse también, como en lingüística, al factor diacrónico y la accidentalidad de la propia evolución. Pese a esta nota escéptica en relación con la capacidad explicativa de la comparación interlingüística, no hay duda de que la adecuación descriptiva de los estudios tipológicos requiere en todo caso elaborar unas muestras comparativas que aspiren a representar equilibradamente, al menos en la medida de lo posible, la diversidad lingüística actual.