EL RITMO INCREMENTAL DE PALABRAS NUEVAS EN LOS REPERTORIOS DE TEXTOS. ESTUDIO EXPERIMENTAL Y COMPARATIVO BASADO EN DOS CORPUS LINGÜÍSTICOS EQUIVALENTES DE CUATRO MILLONES DE PALABRAS, DE LAS LENGUAS INGLESA Y ESPAÑOLA Y EN CINCO AUTORES DE AMBAS LENGUAS

Aquilino Sánchez, Pascual Cantos
Universidad de Murcia

Due to continuous advances in technology and linguistics, computer readable corpora are emerging as one of the most important tools for linguistic studies. Nowadays, computers allow the processing of huge amounts of linguistic data in just splits of seconds, something inconceivable fifteen years ago. However, this has also an inevitable consequence: the more data one processes the more output data one gets. That is, we also need to face how to deal with this increase in output data, which is sometimes too extensive and complex to evaluate properly and in some instances even impossible for an adequate research analysis. There is a need for optimizing this issue, else we shall lose ground to efficiency. One approach for solving this problem is to have at our disposal a predictive tool, that is, a mathematical model, by means of which we could calculate, beforehand, the amount of data —linguistic forms (types) and, probably, lemmas as well— a specific x-word-sized corpus contains. Such a mathematical approach has already been investigated, evaluated and validated for Spanish by means of a Spanish corpus (CUMBRE). This paper attempts at validating the same formula both for English and Spanish, taking as a basis equivalent corpora for each language and some works by outstanding literary writers.

1. El creciente tamaño de los corpus lingüísticos

Hace apenas 20 años casi nadie hablaba de los corpus lingüísticos. En la actualidad, los corpus lingüísticos constituyen uno de los referentes más importantes en multitud de estudios sobre el lenguaje. El índice de la importancia que este instrumento de análisis ha cobrado puede ser ilustrado por ejemplos tan poco usuales hace sólo unos años como el que nos ofrece actualmente la Real Academia Española con la elaboración de un corpus que disfruta de una de las subvenciones más generosas que registra la lingüística española: más de mil millones de pesetas. El proyecto pretende recopilar 100 millones de palabras en un primer estadio y otros 100 millones, de carácter histórico-literario, en una segunda etapa. Pero la cifra de 100 millones de palabras ya ha quedado ampliamente superada por otros proyectos. El corpus COBUILD, pionero en la elaboración de corpus lingüísticos, se acerca o sobrepasa ya los 300 millones de palabras; el proyecto CISLEX, que se está desarrollando en el Centrum für Informations- und Sprachverarbeitung, Instituto perteneciente a la Ludwig-Maximilians-Universität, de Munich, informa que ha procesado más de 1.000 millones de palabras (formas) en lenguas diversas, de las cuales 800 millones se refieren al alemán (provenientes en gran parte de la prensa; Günthner 1996:294).

La tendencia en el crecimiento de los corpus lingüísticos es imparable debido a dos hechos:

(1) el aumento en la capacidad y potencia de los ordenadores, incluidos los PCs, y
(2) la gran cantidad de textos digitalizados a los que es posible tener acceso diariamente.

Ambas razones están cambiando radicalmente el escenario de trabajo relacionado con los corpus lingüísticos. La velocidad del proceso requiere por nuestra parte una rápida y continua adaptación a la nueva realidad.

2. Del millón al billón de palabras

El primer corpus lingüístico, el Brown Corpus, elaborado por dos lingüistas americanos, W. Nelson Francis y Henry Kucera, al inicio de la década de los sesenta, cifró su meta en la recopilación de un millón de palabras del inglés escrito. A imitación de este corpus se recopiló en 1978 el Lancaster-Oslo-Bergen Corpus (LOB), referido al inglés británico. En ambos casos la pretensión había sido lograr un corpus representativo del inglés escrito.

Cuando Sinclair culmina el proyecto COBUILD y este corpus empieza a ser utilizado para estudios lexicográficos y de análisis gramatical diverso, los investigadores constatan que el logro de la representatividad es una tarea más compleja de lo que puede aparentar a primera vista (Sinclair 1991), que un repertorio de un millón de palabras es demasiado limitado para ser considerado 'representativo' de una lengua y que el objetivo de la 'representatividad' debería plantearse sobre bases diferentes, muy especialmente teniendo en cuenta la finalidad para la cual se elabora un corpus y sin perder de vista que el aumento de palabras nuevas (types) no es directamente proporcional al aumento de palabras (tokens) acumuladas (Biber 1993; Sánchez y Cantos, en prensa).

Nuestra propia experiencia en la elaboración de un corpus de español de 20 millones de palabras (Sánchez et al. 1995) y otro paralelo de inglés para fines lexicográficos y de investigación, nos obligaron a recapacitar seriamente sobre las carencias de los corpus reducidos y sobre las dificultades de los corpus de gran tamaño.

Si reparamos en la gran cantidad de datos con que un corpus de 10 o más millones de palabras puede abrumar al investigador (tomando el español como referencia: más de 600.000 ocurrencias de la preposición 'de', más de 40.000 ocurrencias del verbo 'tener' en sus distintas flexiones...), pronto se llega a la conclusión de que en muchos casos y para fines de investigación y analíticos, el ideal sería poder contar con un corpus reducido y representativo. Esto permitiría disminuir la magnitud de los datos que deben analizarse, al mismo tiempo que se conservaría la fiabilidad de tales datos respecto al conjunto que se desea analizar. Aproximarnos a este objetivo, el de la representatividad, implicaría que la muestra o corpus con que trabajemos debería ser capaz de encerrar en sí todos los rasgos esenciales del sistema al cual se refiera la recopilación. En concreto, un corpus representativo de una lengua debería contener los rasgos definitorios de esa lengua, tanto en extensión como en cualidad. Globalmente considerado, este objetivo es difícil de alcanzar en su totalidad. En el trabajo lexicográfico, por ejemplo, un corpus razonablemente representativo, como lo son el corpus CUMBRE o COBUILD (en su formato básico), ofrecen pocas ocurrencias de un buen número de palabras o pocas ocurrencias de las acepciones que integran el significado de muchas voces. Y eso sin tener en cuenta que los ejemplos que aportan son a veces poco adecuados. Es posible argumentar que si una voz o una acepción no

aparece en este tipo de corpus es porque su uso es escaso en la generalidad del uso lingüístico. Pero también es verdad que un diccionario generalista que tiene como objetivo servir a la comunidad de hablantes en su totalidad debe incluir algunas voces y acepciones que no son de uso frecuente en general pero que son fundamentales en el uso de algunas especialidades concretas (medicina, lenguaje, náutica, etc., por ejemplo) y son necesarias siempre que se aborda esa temática. Para lograr ambas finalidades (que aparezcan más voces y más acepciones de cada voz) es necesario ampliar el tamaño del corpus.

La ampliación de los corpus constituye actualmente un problema de menor importancia desde el punto de vista de los utensilios informáticos que deben procesarlo. Un corpus de 10 millones de palabras de texto ocupa unos 70 Mb de espacio en disco y unos 280 Mb en formato indexado. En consonancia con estos datos, un corpus de 50 millones de palabras ocupará unos 350 Mb en versión textual básica y poco más de un Gb en formato indexado. Todo ello es fácilmente asumible por un PC de la familia de los Pentium.

Las dificultades empiezan más bien del lado del investigador. Consideremos algunos datos: si nos ceñimos, por ejemplo, al área de la lexicografía, campo en el cual los corpus lingüísticos han encontrado una de sus primeras aplicaciones prácticas, y nos centramos —una vez más— en el ejemplo del español, constatamos que:

(1) Sobre una forma muy usada, como puede ser el artículo «el», en un extracto del corpus CUMBRE, de 10,3 millones de palabras, esta forma aparece 278.560 veces. En casos como éste, es evidente que la ventaja resulta un inconveniente de cara al manejo de los datos ofrecidos. Hablar de cifras en términos globales es ciertamente impresionante: un corpus de diez millones de palabras generará unas concordancias (palabras con su correspondiente contexto oracional) de, aproximadamente, 200.000 páginas. Sólo las concordancias correspondientes a las ocurrencias de la letra «A» ocuparán unas 6.000 páginas de 50 líneas por página y serán más de 12.000 las que se refieren a las ocurrencias de la preposición «de». La consulta de tales muestras o su análisis implican, por tanto, muchas horas de trabajo, algo que probablemente está vedado al esfuerzo individual y requiere la colaboración y trabajo conjunto de un equipo.

(2) Sin embargo, el número de palabras que aparecen muy frecuentemente en las muestras es más bien reducido. Así, por ejemplo, en el mismo corpus CUMBRE, sobre un total de 9,3 millones de palabras, 800 lemas (elementos equiparables a 'voces de un diccionario') aparecen con una frecuencia superior a 1.000, siendo la más frecuente de 543.044, puesto que ocupa la preposición 'de'. En términos generales, trabajar con muestras que superen las 1.000 frecuencias es trabajoso. Más lo serán aún las 160 voces que superan las 5.000 ocurrencias o las 22 que encabezan la lista de frecuencias, con más de 50.000 casos cada una.

(3) Por otro lado, la mitad de las palabras diferentes que aparecen en un corpus, de cualquier tamaño que éste sea, sólo aparecen una vez. Y es evidente que una sola ocurrencia, que además no ha sido expresamente seleccionada por el investigador y que por tanto puede muy bien ser poco adecuada para los fines específicos del estudioso que la utiliza, tampoco es suficiente.

En conclusión, si nos centramos en la posible utilización de un corpus de unos diez millones de palabras para el trabajo lexicográfico, éste se enfrenta a una riqueza realmente extraordinaria e incluso excesiva en las 1.000 voces más frecuentes, pero a partir del lema 3.500 en secuencia descendente, la frecuencia se limita a 100, caso en el que la utilidad de los ejemplos para el lexicógrafo es ya muy ajustada. En ese mismo corpus, la palabra en la posición 45.000 de frecuencia, aproximadamente, aparece sólo una vez. Y si-

guen todavía otras 45.000 voces que repiten esta misma frecuencia 1. Para tener una idea más exacta de lo que estos datos significan, veamos lo que nos ofrecen algunos diccionarios de español: el diccionario de la RAE contiene unas 83.000 voces, de las cuales un gran número son de uso escaso o nulo. Los diccionarios denominados 'de uso' y de la gama alta o grande, tipo Gran Diccionario de la LE, de SGEL, contiene unas 55.000 voces. En general, este tipo de diccionarios de uso no suele pasar de las 65.000 voces. El lexicógrafo que trabaja con un corpus se enfrenta, pues, a una dificultad doble: por un lado una excesiva abundancia y por otro una notable escasez de ejemplos. La escasez de ejemplos debe solucionarse ampliando el tamaño del corpus, hecho que conlleva necesariamente el incremento desmesurado de un determinado número de voces de frecuencia alta. ¿Cómo hermanar ambos extremos sin disminuir la eficacia del investigador?

3. El cometido del criterio de 'representatividad' de un corpus

El tema de la representatividad de un corpus respecto a una lengua parece que es mejor abordarlo por sectores. Si nos decidimos por la meta de lograr la representatividad total de una lengua, el tamaño del corpus alcanzaría cifras muy altas. Así lo requeriría poder contar con suficiente número de ejemplos útiles en lexicografía, en morfología, en sintaxis o en semántica, por ejemplo. Pero no es nuestra intención entrar en la discusión del concepto de representatividad y de su posible implementación. Solamente nos referimos a ella de manera indirecta, apuntando que para llegar a conclusiones fiables dentro del campo de la representatividad del corpus —sea ésta referida a la totalidad de la lengua o a sectores definidos de ella— es necesario poder precisar cuál sería el mínimo de amplitud de un tal corpus en vistas al logro de los objetivos que nos hayamos propuesto, ya que así sería posible lograr estos objetivos ganando en eficacia e incrementando el ahorro en esfuerzo y en medios económicos. Pues bien, para lograr esta meta es necesario contar previamente con un instrumento de predicción, cual sería, por ejemplo, una fórmula matemática fiable que, a partir de unos determinados supuestos, fuese capaz de proyectar los resultados de manera extensiva.

En el área seleccionada por nosotros en este estudio, —concretada en la predicción del número de palabras y formas (types, tokens)—, una tal fórmula o instrumento debería permitirnos predecir qué número de elementos o formas lingüísticas diferentes contendrá un corpus de una extensión determinada. Contando con tales datos ya sería fácil tomar decisiones sobre la extensión del corpus que precisaríamos para elaborar un diccionario, por ejemplo, de 20.000, 30.000 ó 100.000 voces. Un procedimiento similar sería aplicable en otro tipo de finalidades, cuales podrían ser estudios de orden sintáctico o morfológico.

Para ilustrar la importancia de una tal decisión sobre la extensión de un corpus, considérese la diferencia que supone para un lexicógrafo enfrentarse al análisis de 50, 1.000 ó 50.000 ejemplos a la hora de definir un término. La eficacia y la viabilidad del trabajo aconsejan que los materiales de base se reduzcan a los «necesarios». En caso contrario sería preciso recurrir a una restricción del volumen de datos mediante la aplicación, por ejemplo, de técnicas basadas en el azar. Pues bien, este extremo depende del tamaño del corpus. Si las ocurrencias de la preposición 'de' en español superan las 600.000 en un corpus de 10 millones de palabras, es evidente que la revisión de tal cantidad de frases es inviable para un trabajo eficaz y habría que seleccionar, pongamos por caso, solamente un 5% o un 10% (de 25.000 a 50.000 ocurrencias) para posibilitar el análisis. Si tomáramos en consideración solamente esta preposición y supiéramos con anterioridad que se hace necesario reducir el volumen de datos en una proporción de 1 a 10 ó de 1 a 20 para que aquéllos fuesen manejables por el investigador, entonces no sería necesario recopilar un corpus de 10 millones de palabras, sino que nos bastaría con un corpus de tamaño 5 ó 10

veces más reducido para lograr el volumen de datos que obtendríamos aplicando técnicas de reducción por azar. Tratándose de todo un corpus, esta técnica «reduccionista» afectaría a la globalidad de los términos incluidos en él. Naturalmente, a conclusiones de este tipo solamente se puede llegar una vez que tenemos datos fiables a nuestra disposición. De ahí, una vez más, la necesidad de poder contar con una fórmula predictiva que nos permita tomar decisiones en un estadio previo a la realización.

En nuestro primer trabajo sobre el tema, presentado en el XI Congreso Internacional de AILA, celebrado en Jyväskylä, Finlandia, en agosto de 1996, analizamos con detalle el tema de la predicción de palabras. Los resultados de esta primera investigación aparecerán próximamente publicados en el International Journal of Corpus Linguistics. Como resultado de nuestro estudio, la hipótesis de que era posible contar con un instrumento predictivo respecto al número de palabras diferentes que contendrá un corpus de extensión determinada se confirmó plenamente (Sánchez y Cantos, en prensa). El estudio experimental se basó en un corpus lingüístico del español, de 8 millones de palabras (corpus CUMBRE) y la hipótesis se volvió a confirmar en otros dos sub-corpus temáticos (español periodístico y novela), de menor cuantía.

A raíz de este trabajo intuimos que era muy probable que el procedimiento pudiese aplicarse también a otros idiomas y con resultados similares. Así pues, iniciamos un estudio semejante sobre la predicción de formas lingüísticas en inglés, para luego compararlo con otro corpus equivalente de español. El procedimiento y los resultados se exponen a continuación.

4. Estudio experimental

4.1. Selección de materiales

La fuente para la selección de materiales y textos ha sido para el español el corpus CUMBRE, ya citado (Sánchez et al. 1995), de cuatro millones de palabras, y para el inglés, un extracto de otro corpus paralelo y equivalente en tamaño, que venimos recopilando desde hace algún tiempo con la colaboración de los miembros de nuestro grupo de investigación (EO20-02) en la Universidad de Murcia. En ambos casos nos ajustamos al diseño del corpus explicitado en Sánchez et al. (1995).

Para equilibrar y equiparar ambos corpus en cuanto a contenido y representatividad, nos ceñimos en exclusiva a la modalidad de lenguaje escrito —ya que no disponemos, todavía, de muestras orales suficientes en nuestro corpus de inglés—, y dentro de la modalidad de lenguaje escrito, optamos por variedades lingüísticas tomadas de la prensa y de los libros en general.

Respecto a la extensión de los corpus para el presente estudio, éstos quedaron ajustados y reducidos a un total de 4 millones de palabras cada uno.

Las variedades lingüísticas escritas en ambas lenguas —español e inglés— en cuanto a contenido, temática y porcentajes quedan resumidas en la TAB. 2.

Llevada a cabo esta primera fase de selección y equiparación de las muestras textuales escritas, seguimos el procedimiento aplicado en nuestro trabajo anterior (Sánchez y Cantos, en prensa): dividimos ambos corpus de 4 millones en 16 subcorpus de 250.000 palabras cada uno. A su vez, conservamos porcentajes y proporciones similares en todos y cada uno de los ellos. Idénticos criterios fueron aplicados al contenido y a la temática que subyacen a ambos corpus.

Los subcorpus se ajustan en ambos casos a las siguientes proporciones de muestras textuales escritas (TAB. 1):

TAB. 1. ESTRUCTURA DE LOS SUBCORPUS DE INGLÉS Y DE ESPAÑOL

	ESCRITO		TOTAL ESCRITO
	Libros (palabras)	Prensa (palabras)	
INGLÉS	125.000	125.000	250.000
ESPAÑOL	125.000	125.000	250.000

TAB. 2. CONTENIDOS, TEMÁTICAS Y PORCENTAJES DE LAS MUESTRAS TEXTUALES

LENGUAJE ESCRITO	LIBROS (50%)	Ficción en general (15%)
		Novela histórica (3,75%)
		Novela biográfica (3,75%)
		Novela policíaca/espías (3,75%)
		Novela rosa (3,75%)
		Cuentos (2%)
		Relatos cortos (1,5%)
		Historia (1,5%)
		Economía (1,5%)
		Arte (1,5%)
		Arquitectura (1,5%)
		Técnica/ciencia (1,5%)
		Sociedad (1,5%)
		Psicología (1,5%)
		Filosofía (1,5%)
		Cine (1,5%)
		Deporte (1,5%)
		Viajes (1,5%)
	PRENSA (50%)	Política (7,5%)
		Economía (7,5%)
		Sucesos (7,5%)
		Sociedad (7,5%)
		Editorial (5,5%)
		Cartas al Director (5%)
		Educación (2,5%)
		Deportes (2,5%)
		Clima (1,5)
		Efemérides (1,5%)
		Horóscopo (1,5%)

4.2. Nota preliminar

Conviene aclarar en este punto algunos conceptos habituales en los estudios sobre corpus lingüísticos, a los cuales nos referiremos repetidas veces a lo largo del estudio. Se trata de los términos: palabra o token, forma o type y lema.

Por palabra o token se entiende cualquier ítem lingüístico. Es decir cualquier secuencia grafológica separada de otras por un signo de puntuación o espacio.

El término forma o type hace referencia a las palabras distintas que aparecen en una muestra. Por lo tanto, una palabra que aparezca más de una vez se computará sólo como una forma.

El concepto de lema equivale a lo que en lexicografía se denomina «voz» o « entrada léxica» (es decir, la «palabra» tal y como aparece cuando la buscamos en un diccionario). Por ejemplo, 'llueve' se buscaría en un diccionario bajo la forma 'llover'. Los lemas son, pues, formas no flexionadas o aquellas que, dentro del conjunto de flexiones a las que pueden estar sujetas, se toman como base de las mismas. La secuencia: 'juega, jugaba, juega', estaría, por lo tanto, compuesta por tres palabras o tokens (juega, jugaba, juega), dos formas o types ('juega' y 'jugaba') y un único lema: el verbo 'jugar'.

4.3. Relación entre palabras, formas y lemas en español

Tal y como ha quedado comprobado en nuestro trabajo anterior (Sánchez y Cantos, en prensa), las relaciones entre palabras y formas, por un lado, y palabras y lemas, por otro, constituyen relaciones no-lineales. A más palabras, corresponderán más formas, y a más formas, más lemas. No obstante, puesto que el crecimiento de formas y lemas con respecto al de palabras no es lineal sino curvilíneo (Biber 1993) y, más exactamente parabólico (Sánchez y Cantos, en prensa), el número de nuevas formas y nuevos lemas irá decreciendo y dibujando gráficamente curvas con tendencia a un crecimiento 0, es decir, tenderán a igualarse o a colocarse paralelas con respecto al eje de x (FIG. 1). En nuestro estudio pre-

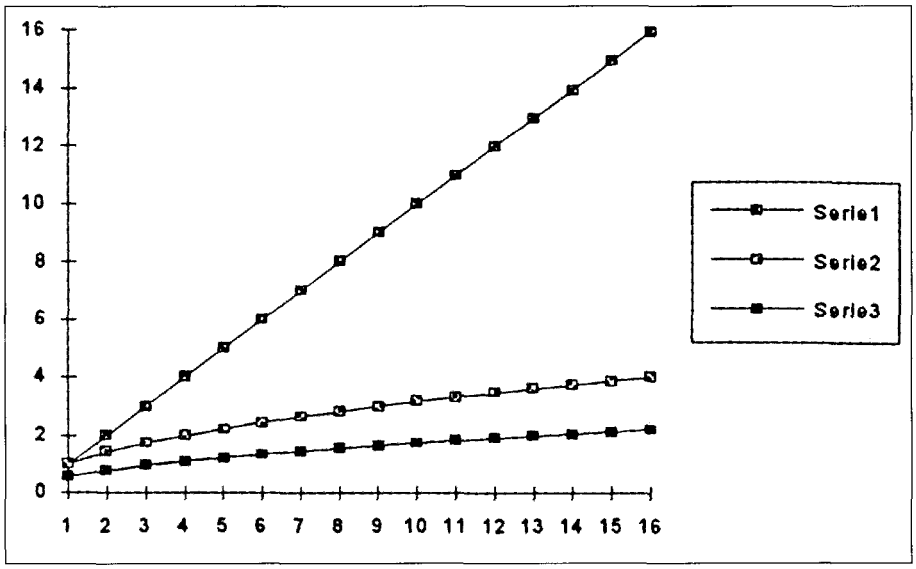

FIG. 1. PALABRAS (SERIE 1), FORMAS (SERIE 2) Y TEMAS (SERIE 3).

vio ya citado, observamos que tanto en el español oral como en el escrito, el crecimiento de las palabras tenía que ser siempre lineal, mientras que el crecimiento de las formas y los lemas dibujaba una trayectoria claramente parabólica. Las relaciones entre palabras y formas, por un lado, y palabras y lemas, por otro reflejaban un comportamiento del tipo $y=\sqrt{x}$.

La observación de este hecho nos llevó a concluir que la curva parabólica descrita no solamente era de tipo curvilíneo, como genéricamente ya habían advertido otros investigadores (Biber 1993), sino que estaba sujeta a patrones matemáticos. Si ello era así, entonces sólo era necesario encontrar la fórmula matemática a la que obedecía. Lograda ésta, la inflexión de la curva podía ser predecible y, en consecuencia, aplicada a los corpus.

Para lograr este objetivo:

(1) determinamos las diversas constantes (K y K_L), para la parábola de las formas y de los lemas. Es decir, calculamos los índices que subyacen en la fijación de la pendiente que dibuja la curva propia de las formas y de los lemas, que es diferente en cada uno de los casos, a pesar de que su tendencia (tendencia a crecimiento 0) sea de la misma índole; y

(2) constatamos que dichas constantes, como índices de los valores absolutos de formas y lemas, aplicados sobre la fórmula base, daban como resultado valores muy cercanos o casi idénticos a los valores reales que resultaban del análisis directo de la realidad del corpus investigado. Tratamos así de verificar la fiabilidad y validez de las fórmulas y sus valores K y K_L.

Nuestra hipótesis inicial quedó plenamente confirmada y constatamos que en español el incremento de formas y lemas seguía unos patrones fijos, perfectamente predecibles a partir de muestras reducidas (de unas 250.000 palabras). Según comprobamos, tales muestras son más que suficientes para:

(1) calcular las constantes K y K_L y

(2) para predecir el número de formas y lemas que una muestra textual contiene o contendrá, con un grado de fiabilidad muy alto (±5%).

De esta manera, no solamente podemos predecir el número de formas y lemas que contiene un texto, sino también, calcular los diversos valores K y K_L para muestras textuales diversas (TAB. 3).

TAB 3. PROYECCIONES DE FORMAS Y LEMAS

	Tokens	K	Types	K_L	Lemas
CUMBRE	10.000.000	53,67	170.668	26,7	84.451
España	10.000.000	52,7	166.652	26,07	82.464
Hispanoam.	10.000.000	55,76	176.328	27,59	87.252
Oral	10.000.000	39,14	123.771	19,36	61.245
Escrito	10.000.000	59,82	189.167	29,6	93.605
Oral (E)	10.000.000	37,96	120.040	18,78	59.399
Oral (H)	10.000.000	41,89	132.467	20,72	65.548
Escrito (E)	10.000.000	59,01	186.606	29,21	92.388
Escrito (H)	10.000.000	61,71	195.144	30,53	96.563

4.4. Hipótesis de trabajo

Partiendo de los datos obtenidos para el español, nuestra hipótesis en el presente estudio queda formulada así: el incremento de formas —partiendo de muestras textuales de 250.000 palabras y conociendo de antemano el número real de formas de esas muestras de 250.000 palabras— es predecible también en corpus textuales de lengua inglesa. Ello significaría que el incremento de formas (y probablemente también de lemas) en inglés seguiría, también, un patrón incremental fijo: una función parabólica del tipo $y=\sqrt{x}$.

4.5. Procedimiento

Para confirmar dicha hipótesis, procedimos de la siguiente manera:

(1) calculamos las formas exactas de cada uno de los tramos de 250.000 palabras en que se subdividieron ambos corpus, mediante un programa de concordancias estándar o concordanciero (tipo OCP, de OUP);

(2) calculamos los valores K de cada uno de esos 16 tramos o subcorpus, para posteriormente determinar la media de todos los valores K;

(3) realizamos una proyección de las formas mediante el valor K medio y comparamos los datos proyectados con los datos reales obtenidos; y

(4) calculamos las diferencias, en cada uno de los tramos, entre los datos reales y las predicciones.

4.6. Resultados

Los resultados obtenidos son los que se exponen en la TAB. 4. El valor K medio resultó ser de 41,88.

Analizando los datos obtenidos para el conjunto de textos ingleses, constatamos que, en términos absolutos, las diferencias entre los datos reales y las proyecciones oscilan entre —1722 formas (con 1.250.000 palabras) y +2812 formas (3.250.000 palabras; ver TAB. 4; columna DIF (r-e). Este hecho queda también patente si observamos el incremento de formas reales en comparación con las formas estimadas (FIG. 2). Los márgenes de error oscilan por lo tanto entre —3,82% y +3,59%; datos bastante optimistas y que a su vez dan sustento a nuestra hipótesis inicial. Debe tenerse en cuenta no solamente que el margen de error es tolerable desde el punto de vista estadístico, sino que la aplicación de nuestra fórmula predictiva es particularmente útil en corpus de gran tamaño y extensión, ya que no tendría sentido realizar cálculos para muestras pequeñas.

TAB. 4. RESULTADOS PARA EL INGLÉS

PALABRAS	FORMAS (reales)	Valores K (reales)	FORMAS (estim.)	DIF (r-e)	DIF % (r-e)
250.000	20.715	41,43	20.940	-225	-1,09%
500.000	29.202	41,29	29.613	-411	-1,41%
750.000	35.974	41,53	36.269	-295	-0,82%
1.000.000	42.130	42,13	41.880	250	0,59%
1.250.000	45.101	40,33	46.823	-1722	-3,82%
1.500.000	50.863	41,52	51.292	-429	-0,84%
1.750.000	55.653	42,06	55.402	251	0,45%
2.000.000	61.079	43,18	59.227	1852	3,03%
2.250.000	62.970	41,98	62.820	150	0,24%
2.500.000	65.190	41,22	66.218	-1028	-1,58%
2.750.000	69.234	41,74	69.450	-216	-0,31%
3.000.000	72.953	42,11	72.538	415	0,57%
3.250.000	78.312	43,43	75.500	2812	3,59%
3.500.000	78.855	42,14	78.350	505	0,64%
3.750.000	81.061	41,85	81.100	-39	-0,05%
4.000.000	84.080	42,04	83.760	320	0,38%

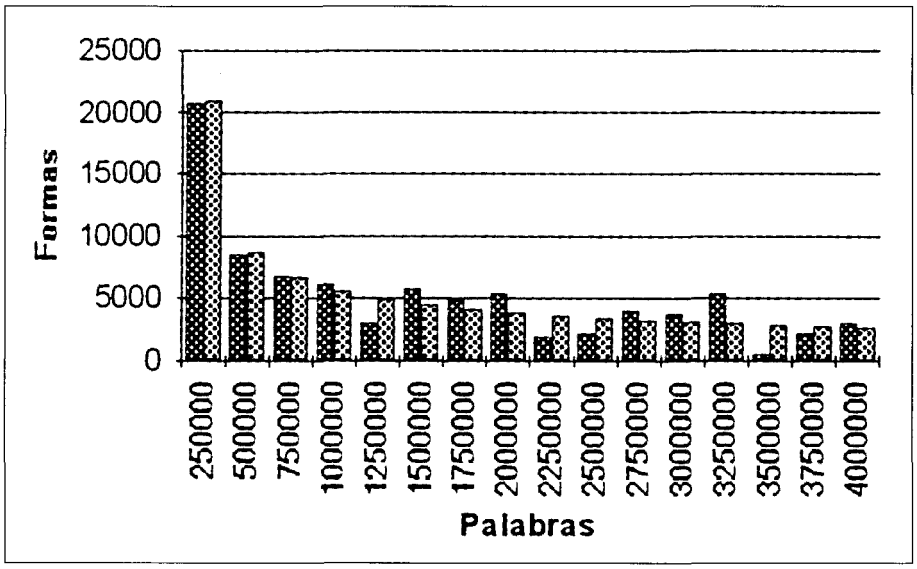

FIG. 2. INCREMENTO DE FORMAS REALES (SERIE 1) VS FORMAS ESTIMADAS (SERIE 2): INGLÉS.

Por otro lado la media porcentual de error es de -0,8%, es decir que la fórmula predictiva calcula una media de 0,8% por encima de las formas que realmente aparecen en las muestras textos. Este hecho sugiere, como veremos más adelante, la aplicación de un índice corrector de la fórmula-TYT: TYPES=K$\sqrt{}$TOKENS.

El resultado obtenido aplicando el procedimiento anteriormente descrito confirma la hipótesis y expectativas que en un principio habíamos expuesto: el comportamiento de las lenguas española e inglesa respecto al incremento de formas lingüísticas en un corpus de x tamaño se atiene a una constante similar que puede ser predicha mediante nuestra fórmula-TYT.

En la FIG. 3 podemos observar las gráficas comparadas relativas al inglés y al español, así como su semejanza incremental. Obsérvese, no obstante, que el punto de inicio difiere, siendo mayor el número de formas en español que en inglés. El hecho es fácilmente explicable porque el español es una lengua más flexionada que el inglés. De ahí que las formas en inglés alcancen la cifra de 84.000, mientras que en español esta cifra asciende a 120.000.

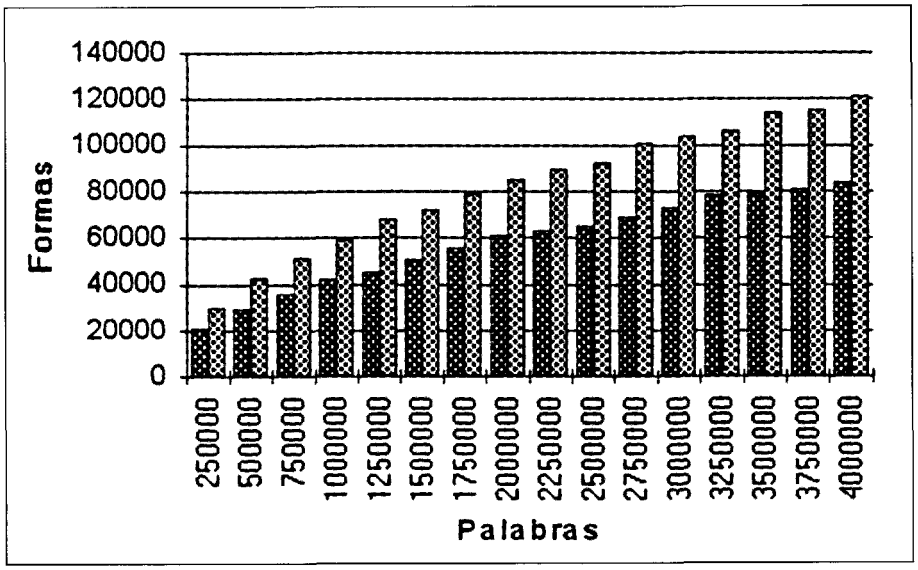

FIG. 3. FORMAS REALES: INGLÉS VS ESPAÑOL.

Este hecho ya se aprecia en el valor K inicial, que en inglés es de 41,3, mientras que en español es de 59,85. Es precisamente este valor inicial el que refleja de entrada la presencia de un mayor número de formas diferentes.

En la FIG. 4 puede apreciarse este factor K en un mismo plano, para facilitar la comparación.

Aplicando la fórmula-TYT, a partir del coeficiente K ya calculado para cada idioma, obtenemos las siguientes gráficas, de crecimiento predecible, para el inglés (FIG. 5) y para el español (FIG. 6).

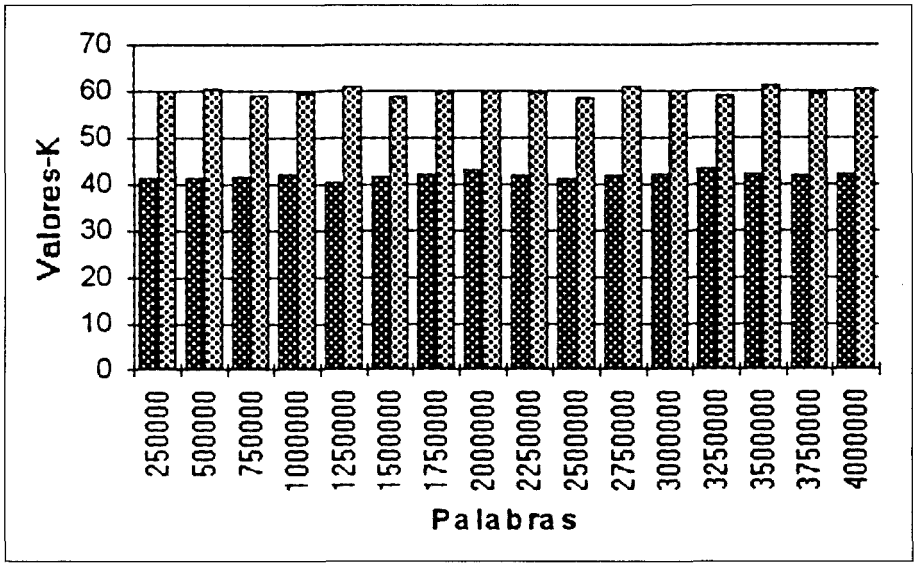

FIG. 4. VALORES-K: INGLÉS VS ESPAÑOL.

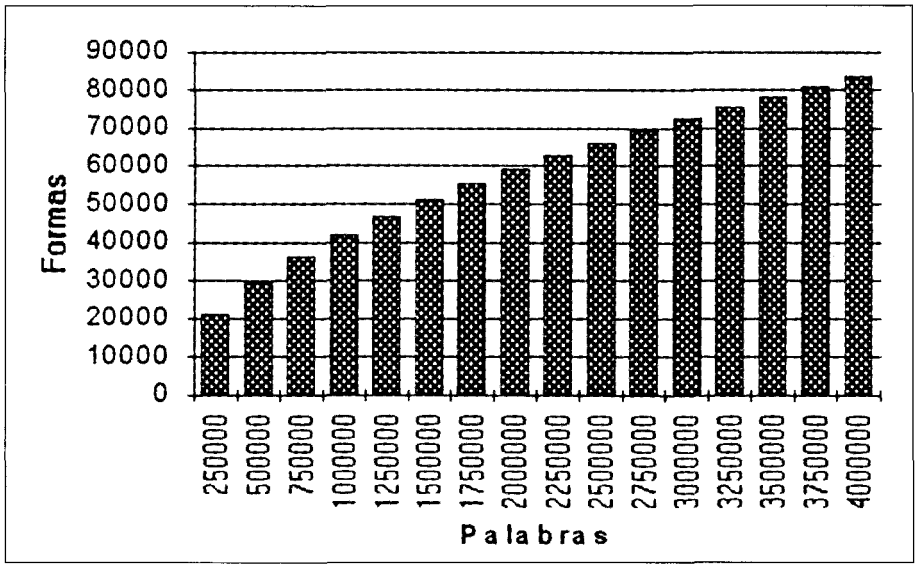

FIG. 5. PREDICCIÓN DE FORMAS: INGLÉS.

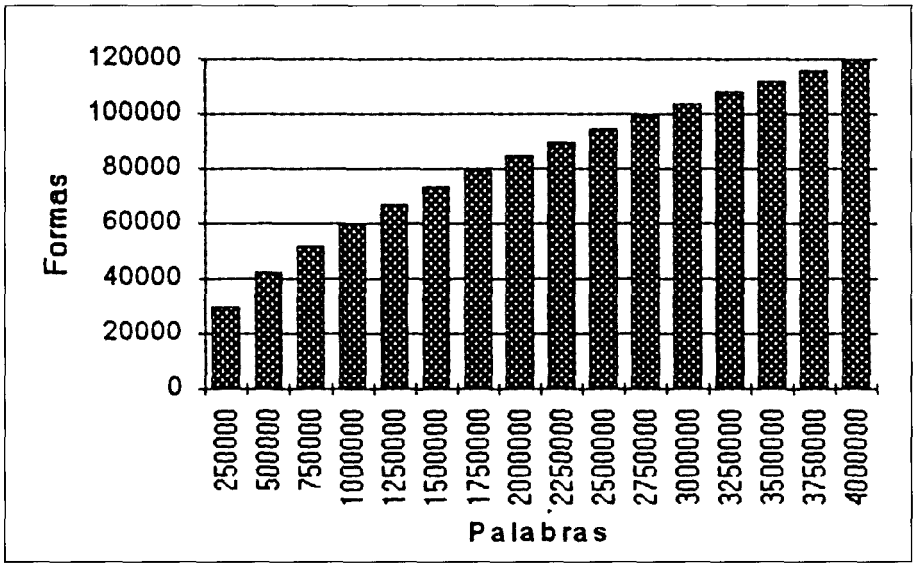

FIG. 6. PREDICCIÓN DE FORMAS: ESPAÑOL.

La gráfica predictiva debe ser comparada con la gráfica resultante de la investigación sobre el número real de formas que se dan en cada uno de los subcorpus. Este extremo es el que se refleja en la FIG. 7 para el inglés, y en la FIG. 8 para el español.

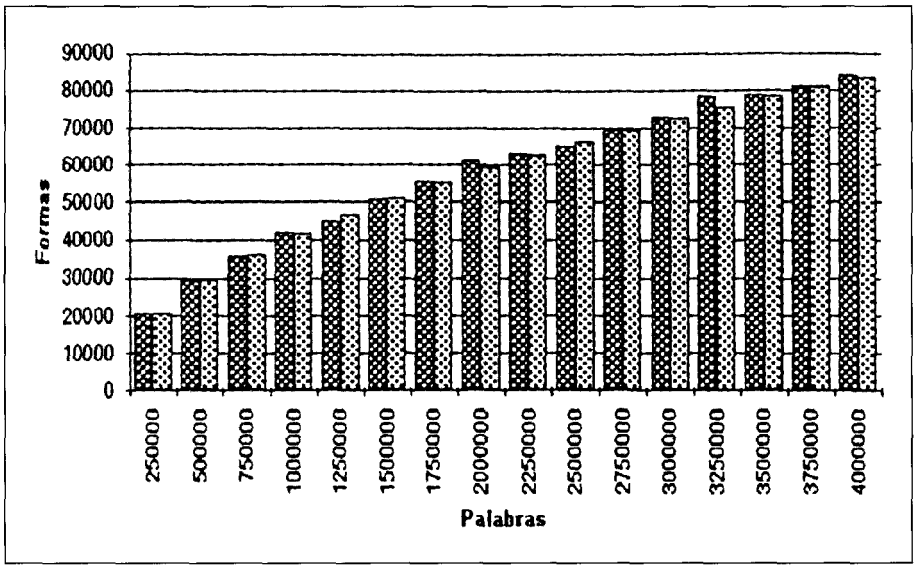

FIG. 7. FORMAS REALES (SERIE 1) VS FORMAS ESTIMADAS (SERIE 2): INGLÉS.

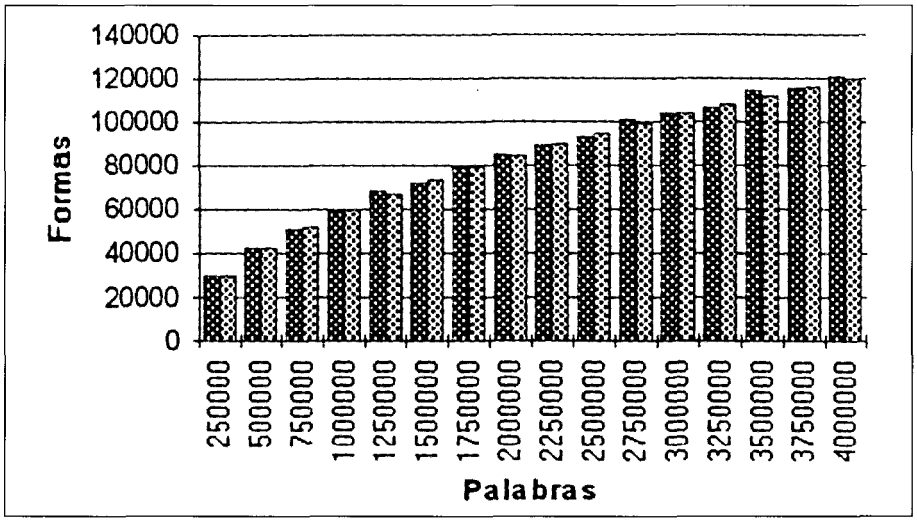

FIG. 8. FORMAS REALES VS FORMAS ESTIMADAS: ESPAÑOL.

En ambos casos queda patente la homogeneidad de incremento tanto en la predicción como en el incremento real, así como las escasas diferencias que se dan entre ambos tipos de datos. Esto nos permite concluir que la fórmula predictiva es fiable y nos permite tomar decisiones basadas en los valores de predicción para un corpus determinado, de una extensión o amplitud determinada.

La confirmación de esta homogeneidad entre datos reales y predictivos, se obtiene también de la comparación de la ratio type-token, tanto en inglés (FIG.9) como en español (FIG. 10).

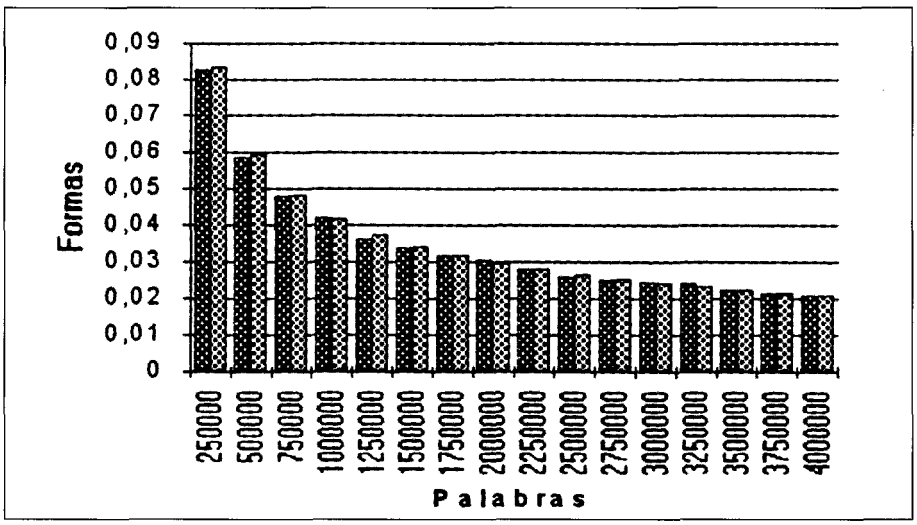

FIG. 9. RATIO TYPE-TOKEN; DATOS REALES VS DATOS ESTIMADOS: INGLÉS.

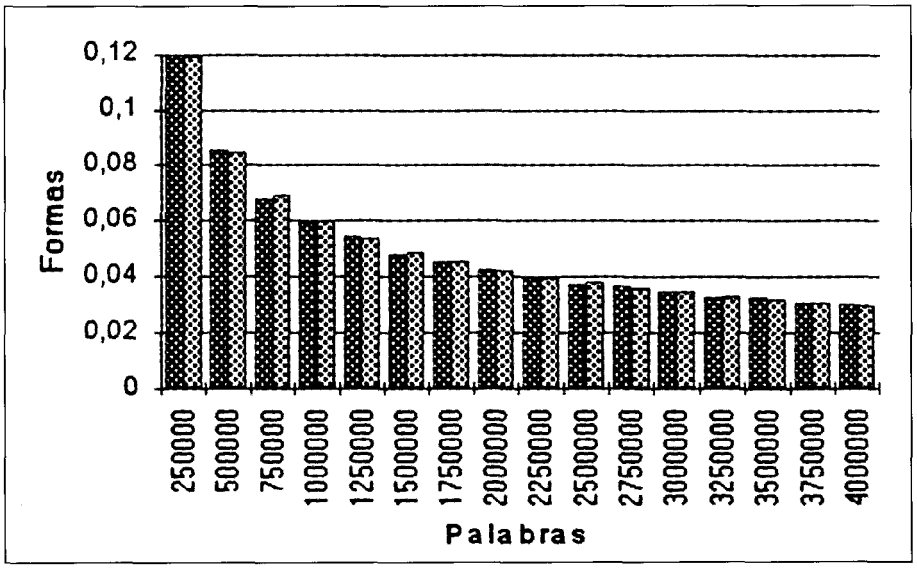

FIG. 10. RATIO TYPE-TOKEN; DATOS REALES VS DATOS ESTIMADOS: ESPAÑOL.

La fórmula predictiva expuesta puede ser utilizada en la elaboración de cuadros o tablas predictivas diversas, hecho que podría ser de gran utilidad para quienes tienen que tomar decisiones sobre el diseño y elaboración de muestras textuales o para el investigador que necesita saber de antemano el número de elementos lingüísticos adecuados para llevar a cabo una determinada investigación. Nuestra fórmula-TYT permite predecir el número de formas no solamente en corpus tipo CUMBRE (de carácter general y razonablemente representativo en su variedad), sino también en cualquier otro tipo de corpus sectorial —prensa, lenguaje técnico, comercial, novela, etc., siempre que se den los requisitos de homogeneidad y coherencia en la recopilación de las muestras. Sobre este tema esperamos ofrecer algunos resultados y tablas predictivas útiles en un próximo futuro.

5. FIABILIDAD DE LA FÓRMULA PREDICTIVA EN VARIOS AUTORES DE OBRAS LITERARIAS: CONRAD, DOYLE, RULFO, CONJUNTO DE NOVELISTAS ARGENTINOS, CERVANTES Y SHAKESPEARE

Nos ha parecido oportuno, con el fin de afianzar la fiabilidad de nuestra fórmula, llevar a cabo una aplicación sectorial sobre la obra de algunos conocidos autores de obras literarias: Conrad, Doyle, Juan Rulfo y un conjunto de seis novelistas argentinos, por un lado y Shakespeare (con cinco de sus obras más importantes) y Cervantes (El Quijote). En parte la elección de estos autores ha sido motivada por el hecho de que sus escritos nos han sido asequibles en forma digitalizada. En cuatro casos podemos considerar que la producción lingüística es finita, puesto que los autores ya han fallecido, y todos ellos gozan en las literaturas de la lengua inglesa y española de una justa y bien merecida fama. Hemos seleccionado algunas obras de cada autor, para luego:

(1) determinar el número de palabras y de formas diferentes con que cuenta una muestra selecta y parcial de tales obras (en el presente caso nos hemos basado en la primera tercera parte de cada una de las obras);

(2) calcular los valores-K reales de cada una de las muestras seleccionadas para cada autor;

(3) comparar los datos reales y estimados del conjunto de obras de cada autor tomadas en consideración en nuestro caso.

Los resultados de estos cálculos y de la consiguiente comparación se reflejan en los cuadros y gráficos siguientes:

TAB. 5. DATOS SOBRE CUATRO OBRAS DE CONRAD

	Tokens	Datos reales		Datos estimados		Diferencias (r-e)	
		Types	K	Types	K	Types	%
Nigger of the 'Narcissus'; Lord Jim; Heart	89.612	10.703	35,75				
of Darkness; y The Secret Agent	271.056	17.795	34,17	18.612	35,75	-817	-4,59

TAB. 6. DATOS SOBRE CUATRO OBRAS DE DOYLE

	Tokens	Datos reales		Datos estimados		Diferencias (r-e)	
		Types	K	Types	K	Types	%
Beyond the City; The Hound of the Baskervilles;	68.378	7.004	26,78				
His Last Bow; y The Casebook of Sherlock Holmes	213.024	12.432	26,93	12.360	26,78	71	0,57

Puede apreciarse con nitidez que los datos reales y los estimados o predecibles mediante la aplicación de nuestra fórmula, son muy similares. Lo cual afianza la fiabilidad del instrumento matemático de predicción.

Además, podemos extraer algunas conclusiones preliminares, si comparamos la relación media entre palabras totales y formas utilizadas por ambos autores ingleses. Observamos, por ejemplo, una mayor riqueza de formas —y en consecuencia una mayor riqueza léxica— en Conrad con respecto a Doyle. El hecho queda reflejado en los valores-K iniciales, que son de 34,17 para Conrad y de 26,93 para Doyle. Un valor-K más alto implica mayor abundancia de formas diferentes. Los resultados obtenidos podrían explicitarse de otra manera más clarificadora: en una hipotética obra de un millón de palabras, Conrad llegaría a utilizar unas 34.170 palabras distintas (types), mientras que Doyle sólo alcanzaría una riqueza léxica de 26.930, es decir casi 8.000 formas distintas menos.

Por otro lado, también puede advertirse que la desviación de los datos estimados frente a los reales es más apreciable en Conrad que en el resto de los autores investigados. El hecho se debe, sin lugar a duda, a las características de la obra de Conrad, quien, como autor no nativo de lengua inglesa, refleja un incremento continuado en el 'aprendizaje' de esta lengua y ello hace que cada obra de temática diferente aporte una mayor cantidad de vocabulario nuevo. Los datos confirman, pues, una característica del estilo de Conrad que reflejan muchos manuales.

En cuanto a Rulfo, los resultados de la investigación son los siguientes:

TAB. 7. DATOS SOBRE CUATRO OBRAS DE RULFO

	Tokens	Datos reales		Datos estimados		Diferencias (r-e)	
		Types	K	Types	K	Types	%
Pedro Páramo y otros relatos cortos	25.807	4.820	30,01				
	74.002	8.495	31,23	8.163	30,01	331	3,9

Nos ha parecido también oportuno corroborar nuestro modelo matemático aplicándolo a dos autores clásicos: Shakespeare y Cervantes. La distancia de tales autores en el tiempo no debería afectar a la validez de la fórmula utilizada, según nuestro criterio.

Seleccionamos cinco obras teatrales de Shakespeare (*Hamlet, Richard III, Othello, Romeo and Juliet* y *Antony and Cleopatra*) y llevamos a cabo los mismos cálculos sobre el total de palabras y formas (tokens/types), siguiendo el mismo procedimiento que el ya descrito para los autores precedentes. Los datos obtenidos son los siguientes:

TAB. 8. DATOS SOBRE CINCO OBRAS DE SHAKESPEARE

	Tokens	Datos reales		Datos estimados		Diferencias (r-e)	
		Types	K	Types	K	Types	%
Hamlet; Richard III; Othello; Romeo and	44.608	5.906	27,96				
Juliet; y Antony and Cleopatra	133.635	10.531	28,8	10.221	27,96	309	2,94

En el caso de Cervantes tomanos como referencia una única obra, *El Quijote*, superior en extensión a las cinco obras de Shakespeare juntas. Comprobamos, en primer lugar, el número total de palabras (185.663) y formas (14.259). Hecho esto, tomamos una muestra de unas 48.000 palabras, calculamos el índice K de tal muestra, proyectamos los datos sobre el total de la obra y comparamos el resultado de tal estimación con los datos reales. Los resultados se ofrecen en la Tab. 9:

TAB. 9. DATOS SOBRE 'EL QUIJOTE'

	Tokens	Datos reales		Datos estimados		Diferencias (r-e)	
		Types	K	Types	K	Types	%
«*Don Quijote de La Mancha*», Parte I	48.431	6.684	33,37				
y Parte II	185.663	14.259	33,09	14.378	33,37	-119	-0,83

Es preciso destacar la homogeneidad de la riqueza léxica en las cuatro obras analizadas de Shakespeare (contrastando ligeramente con lo apreciado en Conrad) y la regularidad de incremento léxico detectado en ambos genios literarios, Shakespeare y Cervantes. La constatación de este hecho nos lleva a pensar que estudios de esta índole pueden ser

válidamente utilizados para confirmar o no determinadas apreciaciones estilísticas, frecuentemente presentes en los manuales de literatura.

Finalmente, quisimos corroborar nuestra modelo matemático con una serie de autores argentinos (Aira, Battista, Borges, Cortázar, Sábato, Saer y Soriano), seleccionando algunos fragmentos de sus obras y siguiendo el mismo procedimiento ya expuesto. Los resultados se ofrecen a continuación:

TAB. 10. DATOS SOBRE ALGUNOS AUTORES ARGENTINOS

	Tokens	Datos reales		Datos estimados		Diferencias (r-e)	
		Types	K	Types	K	Types	%
Aira, Battista, Borges, Cortázar, Sábato,	19.112	5.499	39,78				
Saer y Soriano	55.648	9.865	41,82	9.384	39,78	481	4,87

Como era de esperar, la riqueza léxica de estos seis autores argentinos es mayor tomada en su conjunto que individualmente. Este hecho se manifiesta en el factor K, sensiblemente superior (39,78 / 41,81) al que encontramos en los autores individuales anteriormente estudiados (entre 25 y 30). Pero la predicción fundamentada en nuestra fórmula sigue siendo igualmente válida en todos sus términos, como reflejan los valores reales contrastados con los estimados y la escasa diferencia que media entre ellos.

6. Conclusiones

Los resultados obtenidos en todas las muestras analizadas, tanto en corpus generales como en las obras parciales de algunos autores literarios, no dejan lugar a dudas: podemos concluir que no solamente los textos en español —como quedó patente en nuestro anterior estudio (Sánchez y Cantos, en prensa)—, sino también los textos en lengua inglesa siguen unos patrones fijos respecto al incremento de formas nuevas. El estudio empírico se ha limitado al lenguaje escrito (libros y prensa). Pero no es aventurado predecir que el lenguaje oral también se atiene, muy probablemente, a un patrón incremental parabólico del tipo $y=\sqrt{x}$, como ya quedó patente en el estudio realizado sobre el español.

Así pues, mediante la aplicación de nuestra fórmula podemos predecir que en cada una de las lenguas estudiadas, el número de formas lingüísticas que cabe esperar de un corpus de textos escritos, de estructura equivalente a la muestra analizada y de extensión como la especificada en la columna de la izquierda, será el siguiente:

TAB. 11. PROYECCIÓN DE FORMAS EN ESPAÑOL E INGLÉS

	ESPAÑOL		INGLÉS	
Tokens	Types	Ratio T-T	Types	Ratio T-T
10.000.000	189.167	0,0189167	132.436	0,0132436
20.000.000	267.523	0,0133762	187.293	0,0093647
30.000.000	327.648	0,0109216	229.386	0,0076462
40.000.000	378.335	0,0094584	264.872	0,0066218
50.000.000	422.991	0,0084598	296.136	0,0059227
60.000.000	463.364	0,0077227	324.401	0,0054067
70.000.000	500.490	0,0071499	350.393	0,0050056
80.000.000	535.046	0,0066881	374.586	0,0046823
90.000.000	567.502	0,0063056	397.309	0,0044145
100.000.000	598.200	0,0059820	418.800	0,0041880

REFERENCIAS

Biber, D. 1993. Representativeness in Corpus Design. Literary and Linguistic Computing 8 (4): 243-257.

Günthner, F. 1996. Electronic Lexica and Corpora Research at CIS. *International Journal of Corpus Linguistics*, I/2, 1996, 287-301.

Sánchez, A. y P. Cantos 1996. Predictability and Representativeness of Words, Word Forms and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus -an 8 Million Word Corpus of Contemporary Spanish. Comunicación presentada en el XI Congreso Internacional de Lingüística Aplicada (AILA), Universidad de Jyväskylä (Finlandia; agosto 1996).

Sánchez, A. y P. Cantos (en prensa). Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus -an 8 Million Word Corpus of Contemporary Spanish. *International Journal of Corpus Linguistics.*

Sánchez, A. et al. 1995. *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones.* Madrid: Sociedad General Española de Librería.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

INDICE