CANELA. Confederación Académica Nipona, Española y Latinoamericana, Universidad Doshisha, Kioto, 26/mayo/2018

Métodos de lingüística española de corpus en línea

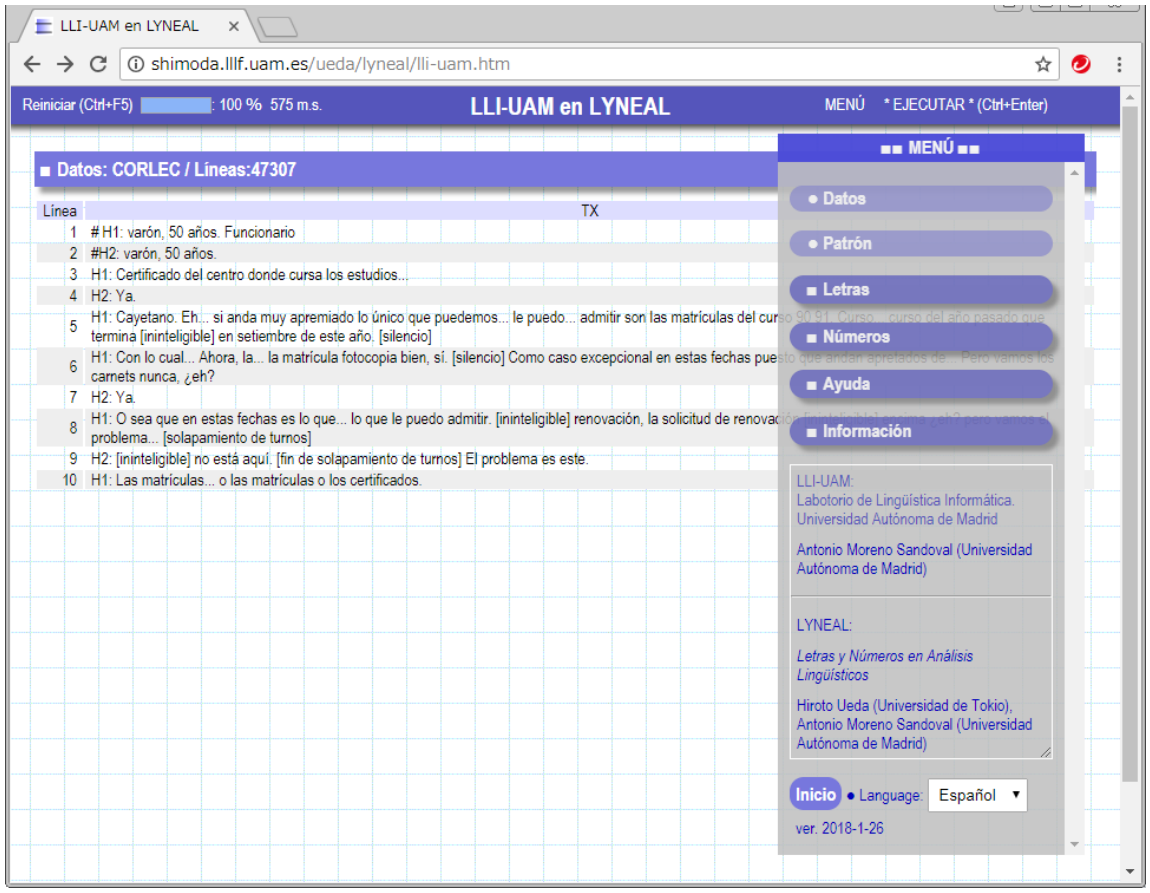
Medidas de diversidad léxica con evidencias documentales

Hiroto Ueda, Universidad de Tokio
Antonio Moreno Sandoval, Universidad Autónoma de Madrid

Material: Obras de Benito Pérez Galdós (Proyecto Gutenberg)

Marianela (1878), *Trafalgar* (1873), *Misericordia* (1897)
y *Fortunata y Jacinta* (Fortunata: 1886)
http://www.gutenberg.org/
[26 de mayo, 2018]

Herramienta: nuestro sistema LYNEAL
«*Letras y Números en Análisis Lingüísticos*»
http://shimoda.lllf.uam.es/ueda/lyneal/
https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/index.html
[26 de mayo, 2018]



Sistema LYNEAL «*Letras y Números en Análisis Lingüísticos*»

Google: lyneal

1. Medidas de diversidad léxica

1. 1. Tipo - Total y Forma - Lema

Adivinanza-1: 'Nieve'

La sábana de Leonor todo lo tapa y el río no. Nieve.

Núm.	Contexto anterior (*)	**Forma**	Contexto posterior (=)
3	La sábana	**de**	Leonor todo lo tapa y el río no. Nieve
9	La sábana de Leonor	**el**	río no. Nieve

	todo lo tapa y		
1		**La**	sábana de Leonor todo lo tapa y el río no. Nieve
4	La sábana de	**Leonor**	todo lo tapa y el río no. Nieve
6	La sábana de Leonor todo	**lo**	tapa y el río no. Nieve
12	La sábana de Leonor todo lo tapa y el río no.	**Nieve**	
11	La sábana de Leonor todo lo tapa y el río	**no**	. Nieve
10	La sábana de Leonor	**río**	no. Nieve

	todo lo tapa y el		
2	La	**sábana**	de Leonor todo lo tapa y el río no. Nieve
7	La sábana de Leonor todo lo	**tapa**	y el río no. Nieve
5	La sábana de Leonor	**todo**	lo tapa y el río no. Nieve
8	La sábana de Leonor todo lo tapa	**y**	el río no. Nieve

Tabla 1.1.a. Concordancia de *Adivinanza. Nieve*.

Forma	de	el	La	Leonor	lo	Nieve	no	río	sábana	tapa	todo	y
Frecuencia	1	1	1	1	1	1	1	1	1	1	1	1

Tabla 1.1.b. Frecuencia de formas de *Adivinanza. Nieve*.

Tipo = 12 / Total = 12

Adivinanza-2: 'Trampa'

En el huerto hay un muerto. El muerto tiene un cautivo. Llega un vivo, toca al cautivo y el muerto mata al vivo. Trampa.

Núm.	Contexto anterior (*)	**Forma**	Contexto posterior (=)
16	En el huerto hay un muerto. El muerto tiene un cautivo.	**al**	cautivo y el muerto mata al vivo.

	Llega un vivo, toca		Trampa
22	l huerto hay un muerto. El muerto tiene un cautivo. Llega un vivo, toca al cautivo y el muerto mata	**al**	vivo. Trampa
11	En el huerto hay un muerto. El muerto tiene un	**cautivo**	. Llega un vivo, toca al cautivo y el muerto mata al vivo. Trampa
...	...	...	...
23	* hay un muerto. El muerto tiene un cautivo. Llega un	**vivo**	. Trampa

	vivo, toca al cautivo y el muerto mata al		
18	En el huerto hay un muerto. El muerto tiene un cautivo. Llega un vivo, toca al cautivo	**y**	el muerto mata al vivo. Trampa

Tabla 1.1.c. Concordancia de *Adivinanza.Trampa.*

F	al	cautivo	el	en	hay	huerto	llega	mata	muerto	tiene	toca	trampa	un	vivo	y
Fr.	2	2	3	1	1	1	1	1	3	1	1	1	3	2	1

Tabla 1.1.d. Frecuencia de formas de *Adivinanza. Trampa*.

Tipo = 15 / Total = 18

Lema y Forma

La sábana de Leonor todo lo tapa y el río no. Nieve.

Formas 'La' / 'el' → Lema EL
Forma 'tapa' → Lema TAPAR.

Tanto Lema como Forma pueden ser calculados en Tipos y/o Total de Figuración.

1. 2. Ratio de Tipo por Total (RTT)

«Type Token Ratio» (TTR)

RTT = Tipo / Total = 15 / 18 = 0.833

El inconveniente de la RTT consiste en que el Tipo no crece al mismo modo que el Total, puesto que al aumentar el Total de palabras en distintos textos, se repiten cada vez más palabras y, por consiguiente, no aumenta el número de Tipo en la misma proporción (sec. 1.4.), dando como resultado una RTT cada vez más reducida. cf. Baker et al. (2006: 159)

Cap.	22	10	13	16	12	18	15	1	6	3	5
Total	844	1299	1382	1392	1434	1638	1729	1786	2032	2066	2171
Tipo	453	614	576	627	583	675	678	788	718	813	987
RTT	0.537	0.473	0.417	0.45	0.407	0.412	0.392	0.441	0.353	0.394	0.455

Cap.	7	17	11	20	8	2	14	4	9	19	21
Total	2206	2472	2493	2557	2615	2653	2684	3262	3666	3749	4826
Tipo	849	939	975	921	943	1017	1032	1237	1336	1163	1492
RTT	0.385	0.38	0.391	0.36	0.361	0.383	0.385	0.379	0.364	0.31	0.309

Tabla 1.2.a. Total, Tipo y Ratio de Tipo / Total (RTT) de *Marianela*

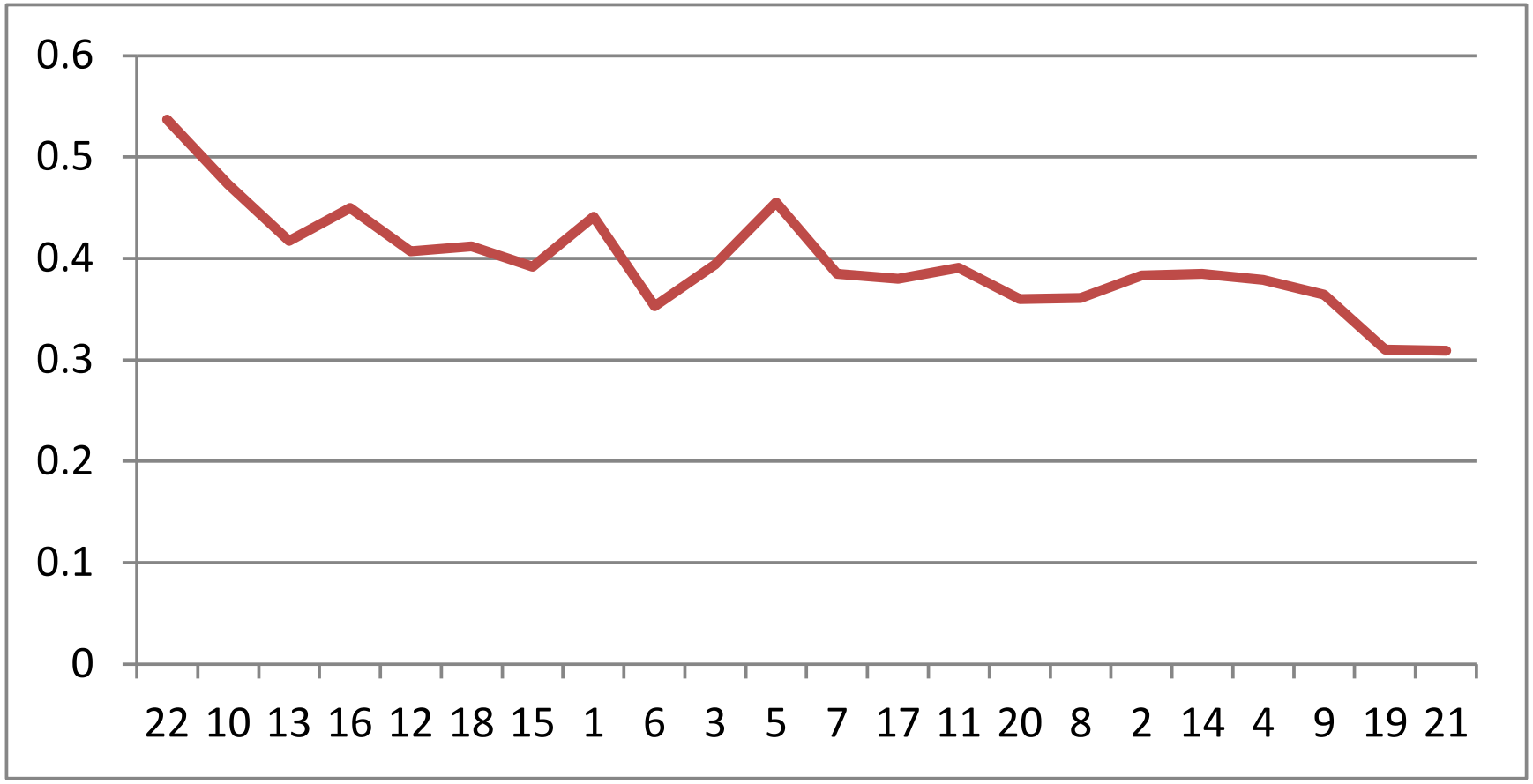


Fig. 1.2. Ratio de Tipo / Total (RTT) de *Marianela*

1. 3. Índice de Tipo (IT)

Zipf (1936: 44-48, 1949: 22-27):

R(ango) * F(recuencia) = C(onstante)

R	F = C / R	Int(F)	R * F = C
R(1): 1	F(1) = M: 50	50	50
2	25.00	25	50
3	16.67	16	50
4	12.50	12	50
5	10.00	10	50
…	…	…	…
49	1.02	1	50
R(n): 50	F(n): 1.00	1	50
51	0.98	0	50

Tabla 1.3. Rango (R), Frecuencia (F) y Constante (C) en la fórmula de Zipf (1)

$F(1) = M = R(n) = \text{Tipo} = C$

Demostración:

$R * F = C$ ← Fórmula de Zipf

$R(1) * F(1) = C$ ← R(1): rango 1, F(1)

$1 * F1 = C$ ← R(1) = 1

$1 * M = C$ ← F(1) = M: máxima frecuencia

$M = C$ ← Máxima frecuencia = constante

$F = C / R$ ← Fórmula de Zipf: R * F = C

$F = M / R$ ← M = C

$F = M / R \geq 1$ ← La frecuencia debe ser igual o más de 1.

$M \geq R$ ← Multiplicar ambos lados por R, que es positivo

$R \leq M$ ← Significa que el rango es igual o menos de M.

$R(n) = M$ ← Significa que el último rango = frecuencia máxima.

$Tipo = M$ ← R(n) = Tipo: número de tipos

$Tipo.z. = M$

Índice de Tipo (IT):

$$IT = Tipo / (Tipo + Tipo.z.) = Tipo / (Tipo + M)$$

$IT \rightarrow 0$ cuando $Tipo \rightarrow 0$

$IT = 0.5$ cuando $Tipo = Tipo.z$

$IT \rightarrow 1$ cuando $Tipo \rightarrow \infty$ y/o $Tipo.z. \rightarrow 0$ ($M \rightarrow 0$)

1. 4. Índice de Hápax (IH)

R	F = C / R	Int(F)	R * F = C
R(1): 1	F(1)=M: 50	50	50
R(2): 2	25.00	25	50
R(3): 3	16.67	16	50
…	…	…	…
R(24): 24	2.08	2	50
R(25): 25	**F25: 2.00**	**2**	**50**
R(26): 26	1.92	1	50
R(27): 27	1.85	1	50
…	…	…	…

R(n): 50	Fn: 1.00	1	50
R(51): 51	0.98	0	50

Tabla 1.4. Rango (R), Frecuencia (F) y Constante (C) en la fórmula de Zipf (2)

$$Hápax.z = M / 2$$

Demostración:

$F = C / R$ ← Fórmula de Zipf: $R * F = C$

$F = M / R$ ← $M = C$ (demostrado anteriormente)

$F = M / R \geq 2$

← La frecuencia de no hápax debe ser igual o más de 2.

$M \geq 2\ R$

← Multiplicamos ambos lados por R, que es positivo.

$R \leq M / 2$

← Dividimos ambos lados por 2 e intercambiamos los dos lados.

$Hápax.z. = Rn - M / 2 = M - M / 2 = M / 2$

← Hápax.z. debe ser fuera del ámbito de $R \leq M / 2$

$Hápax.z = M / 2$

«Índice de Hápax» (IH):

IH = Hápax / (Hápax + Hápax.z.) = Hápax / (Hápax + M / 2)

IH → 0 cuando Hápax → 0
IH = 0.5 cuando Hápax = Hápax.z.
IH → 1 cuando Hápax → ∞ y/o Hápax.z.→0 (M→0)

2. Aplicaciones

2. 1. Total - Tipo - Hápax - Máximo

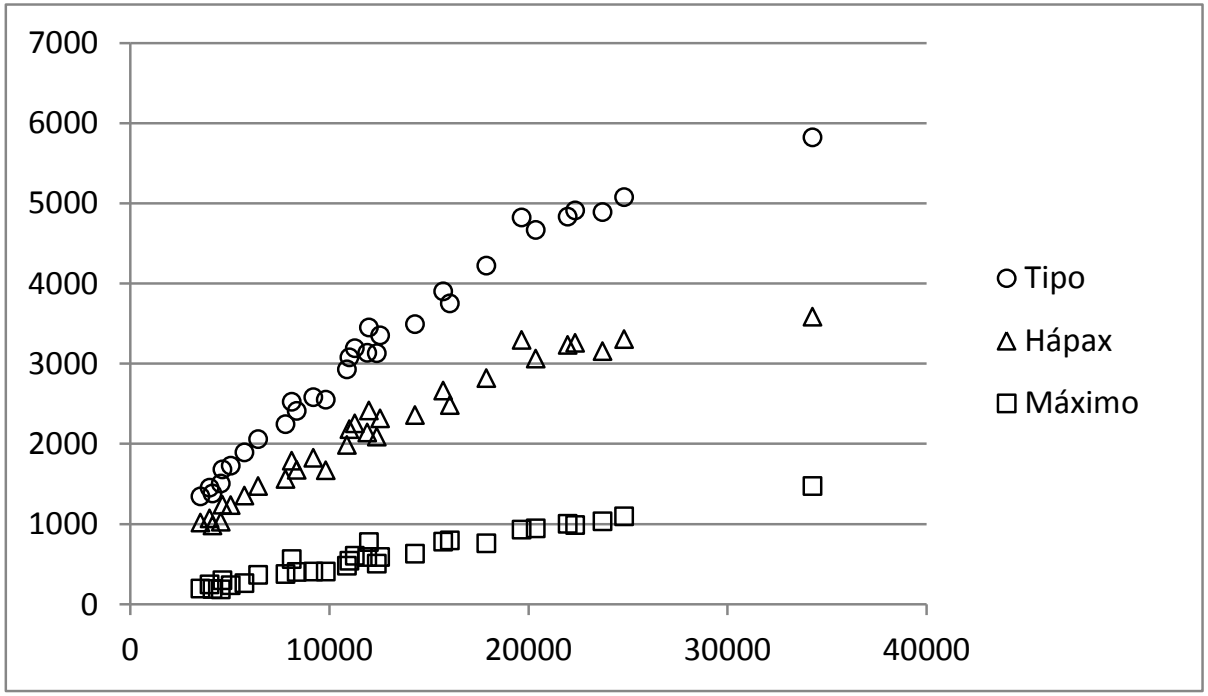
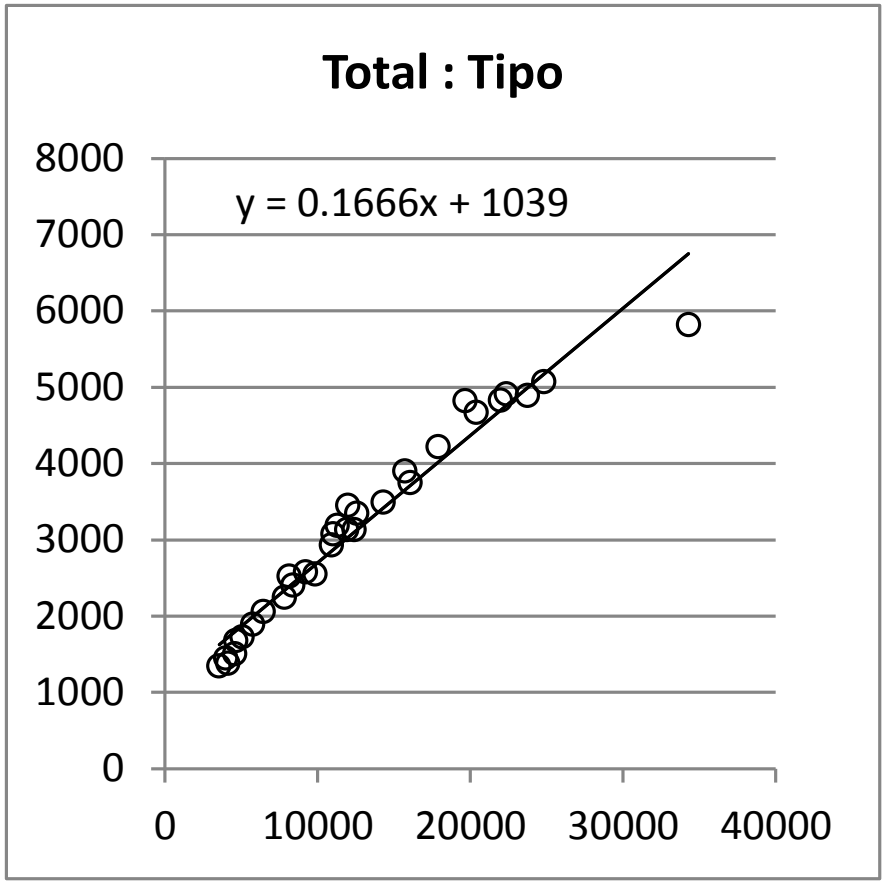


Fig. 2.1.a. Total y Tipo - Hápax - Máximo en *Fortunata y Jacinta*



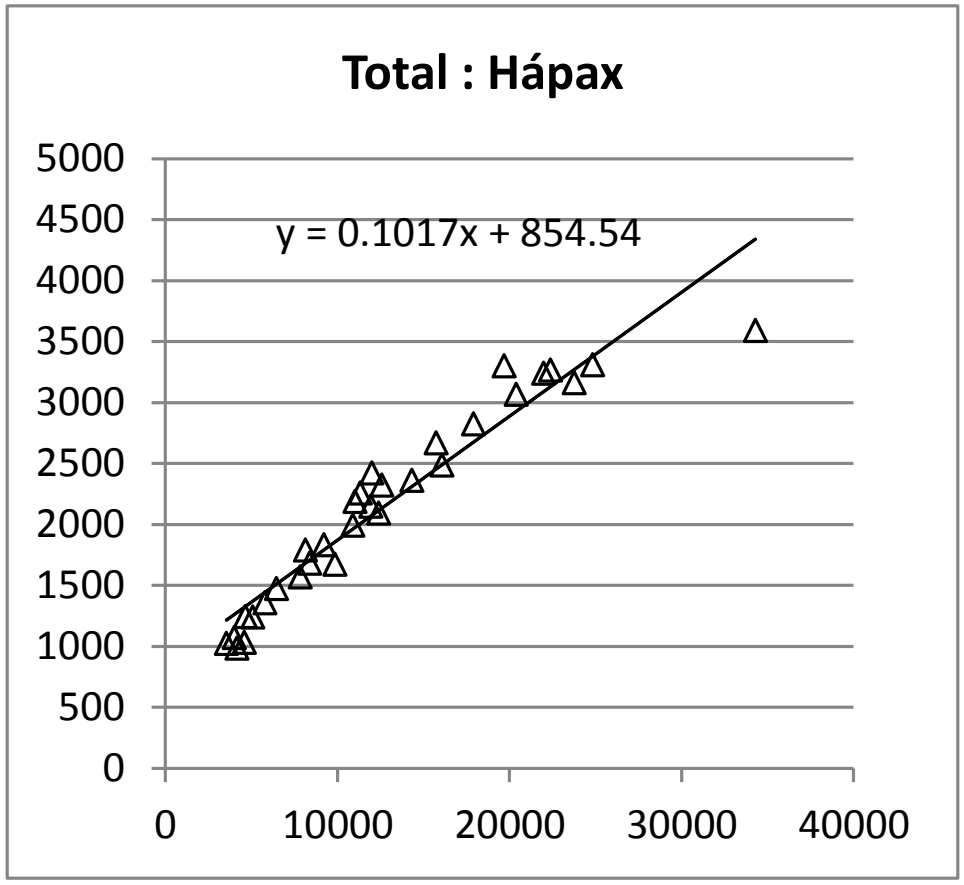
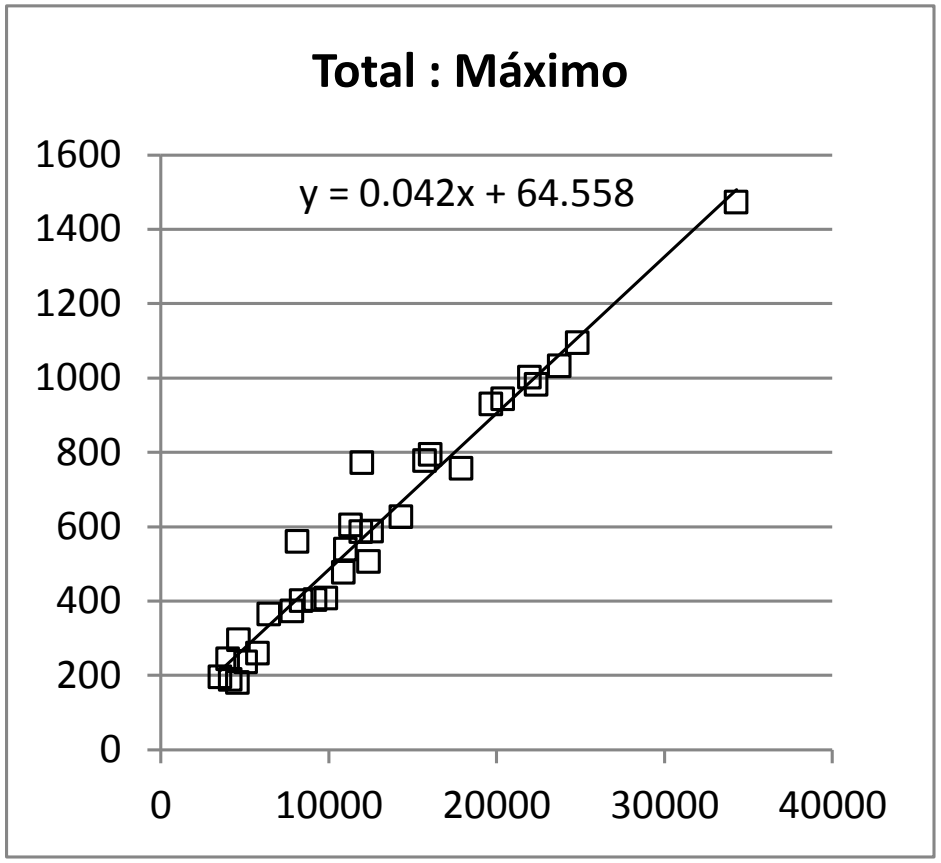


Fig. 2.1.b. Total y Tipo (R. = .978) / Fig. 2.1.c. Total y Hápax (R.= 959)



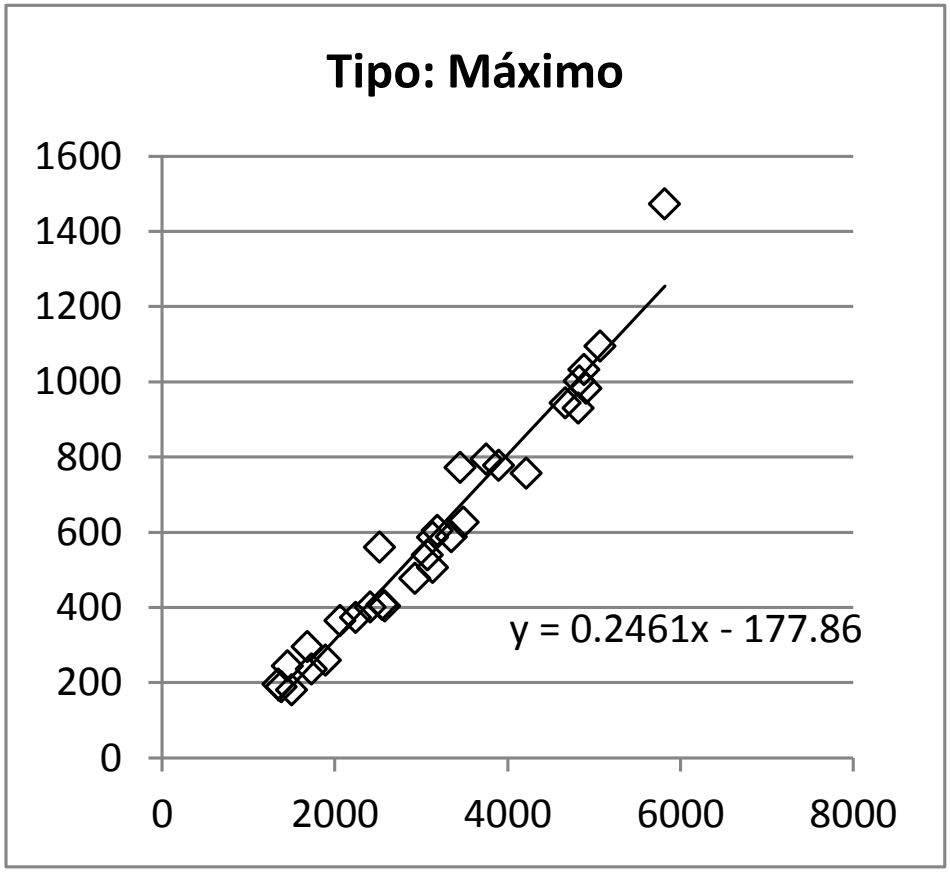


Fig. 2.1.d. Total y Máximo (R. = .981) / Fig. 2.1.e. Tipo y Máximo (R. = .978)

2. 2. Curva descendente de RTT

Trafalgar. Capítulo	3	2	7	10	17	1	5	6	16
Total	1 300	1 402	2 000	2 183	2 298	2 503	2 591	2 723	2 777
Tipo	609	646	836	910	896	1 040	1 033	1 063	1 030
Máximo	78	65	81	103	116	168	125	144	144
Hápax	472	494	633	663	636	788	768	767	732
Tipo / Total (RTT)	0.468	0.461	0.418	0.417	0.39	0.416	0.399	0.39	0.371
Índice de Tipo	0.886	0.909	0.912	0.898	0.885	0.861	0.892	0.881	0.877
Índice de Hápax	0.924	0.938	0.94	0.928	0.916	0.904	0.925	0.914	0.910

Trafalgar. capítulo	9	14	11	13	12	4	8	15
Total	3 016	3 184	3 482	3 837	4 140	4 200	4 579	4 938
Tipo	1 245	1 195	1 255	1 355	1 450	1 470	1 671	1 603
Máximo	164	139	194	183	209	200	235	220
Hápax	936	844	892	928	1 032	1 041	1 225	1 090
Tipo / Total	0.413	0.375	0.360	0.353	0.350	0.350	0.365	0.325
Índice de Tipo	0.884	0.896	0.866	0.881	0.874	0.880	0.877	0.879
Índice de Hápax	0.919	0.924	0.902	0.910	0.908	0.912	0.912	0.908

Tabla 2.2.a. Total, Tipo, Máximo, Índice de Tipo, Índice de Hápax en *Trafalgar*

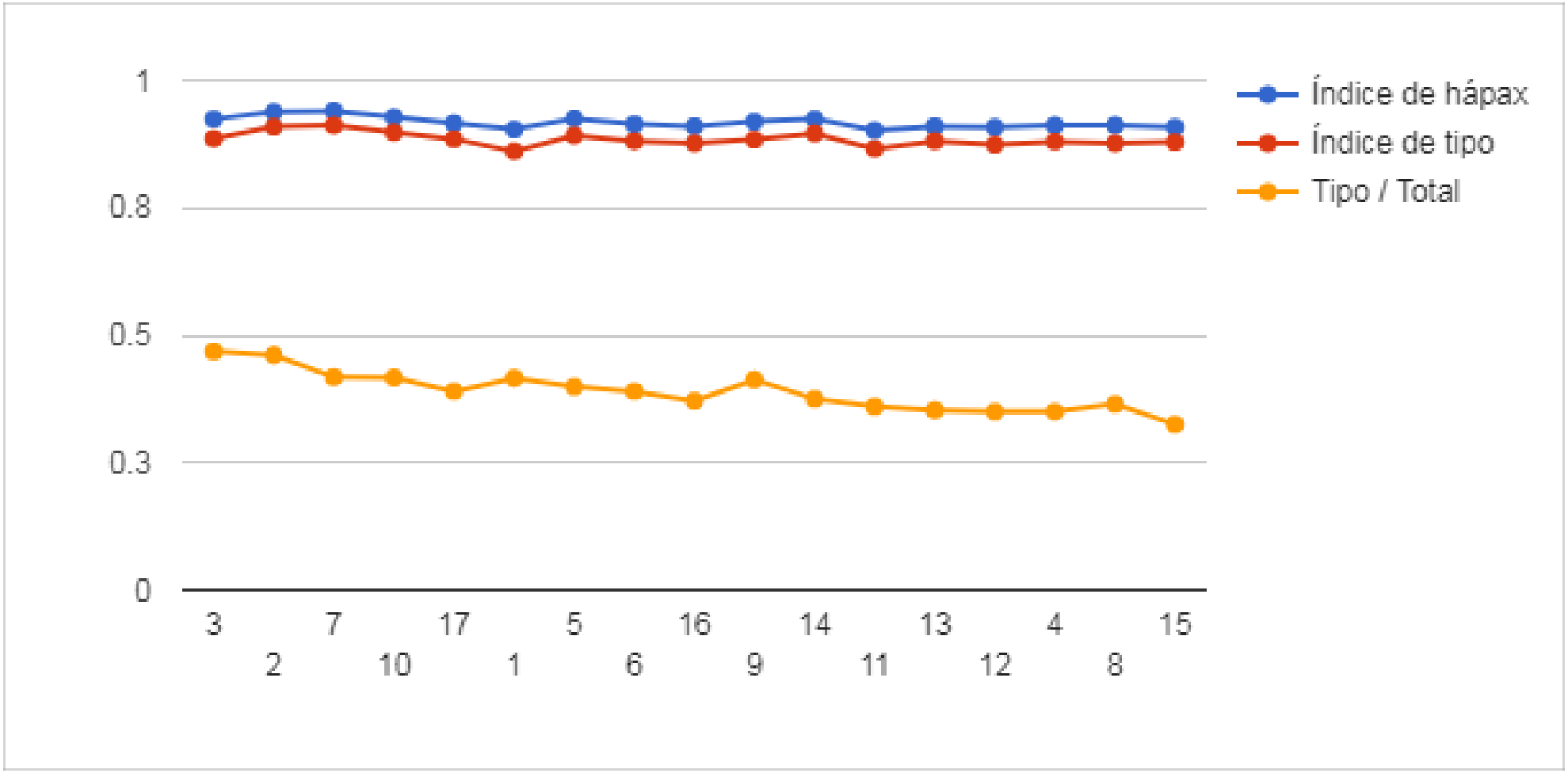


Figura 2.2.a. IT, IH, RTT, RRTT en *Trafalgar*

Obra	Marianela	Trafalgar	Misericordia	Fortunata y Jacinta
Total	50 960	51 157	83 840	394 606
Tipo	8 200	8 443	12 764	29 368
Máximo	2 493	2 558	4 270	18 247
Hápax	4 872	5 011	7 566	14 398
Tipo / Total	.161	.165	.152	.074
Índice de tipo	.767	.767	.749	.617
Índice de hápax	.796	.797	.780	.612

Tabla 2.2.b. Total, Tipo, Máximo, Índice de Tipo, Índice de Hápax en 4 obras de Pérez Galdós

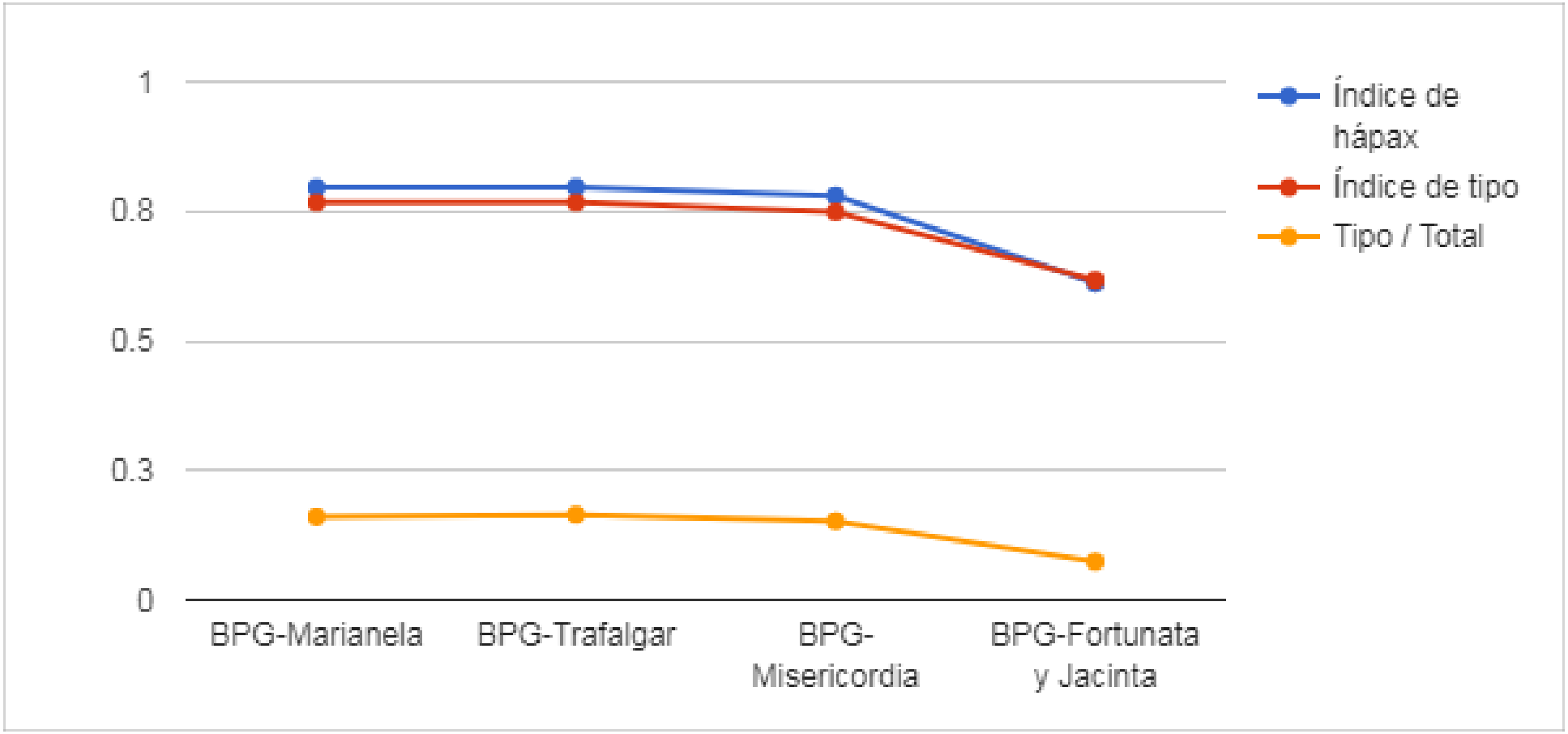


Fig 2.2.b. Índice de Hápax , Índice de Tipo, Ratio de Tipo / Total en 4 obras de Pérez Galdós

2. 3. Formas y lemas

Diversidad de frecuencia	Fortunata	Fortunata-lema
Total	394 606	393 797
Tipo	29 368	13 857
Máximo	18 247	31 362
Hápax	14 398	4 962
Tipo / Total	0.074	0.035
IT	0.617	0.306
IH	0.612	0.240

Tabla 2.3. Diversidad léxica de *Fortunata y Jacinta* en formas y lemas

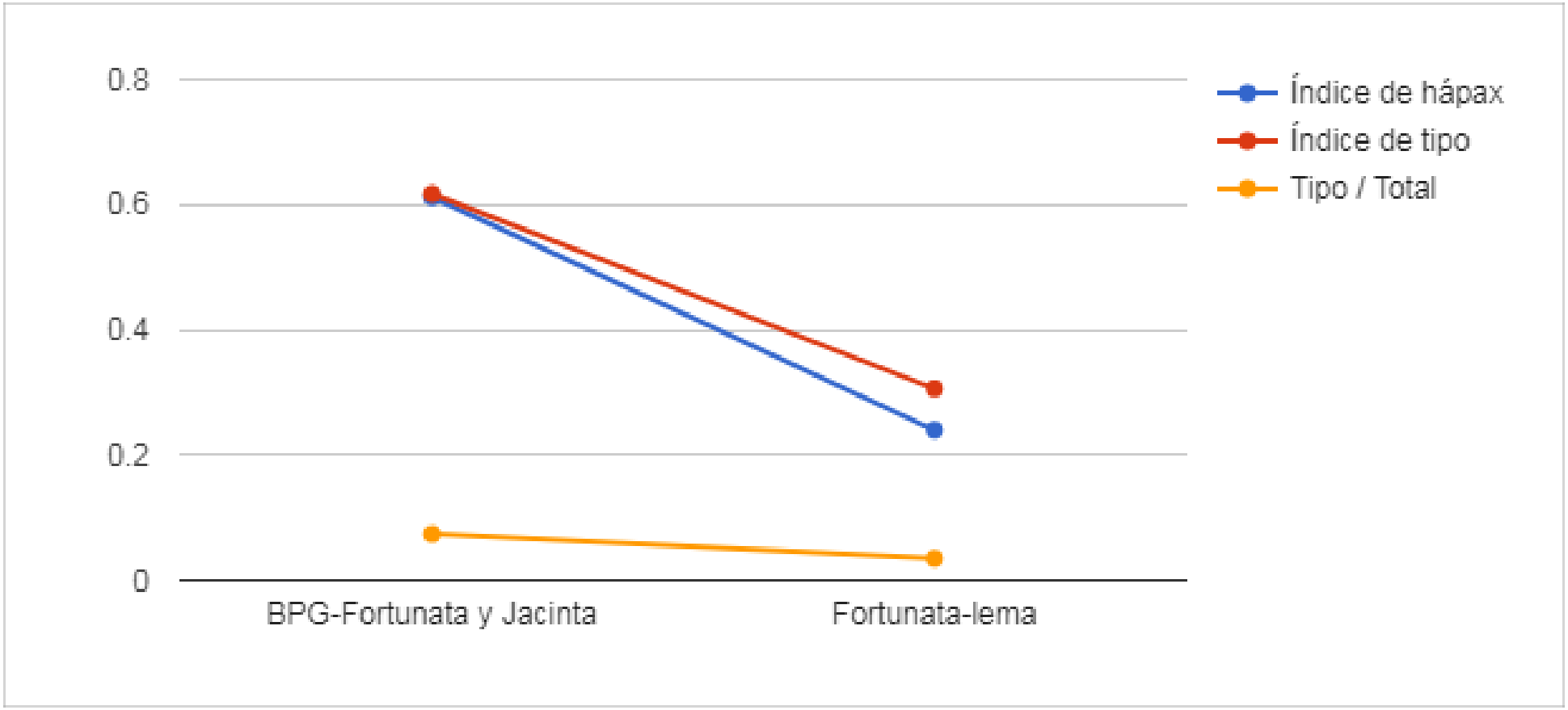


Fig. 2.3. Diversidad léxica de formas y lemas en *Fortunata y Jacinta*

2. 4. Nuevas formas y nuevos lemas

Baker et al. (2006: 159):

> (...) high frequency words like *the* tend to be repeated whereas the probability of new types of words appearing will always decrease, the larger the corpus size. Therefore, the type/token ratio tend to reveal more about corpus size than lexical repatition or uniqueness.

(...) las palabras de alta frecuencia como *the* tienden a repetirse, mientras que la probabilidad de que aparezcan nuevos tipos de palabras siempre disminuirá, cuanto mayor sea el tamaño del corpus. Por lo tanto, la ratio tipo / token tiende a revelar más el tamaño del corpus que la repartición o la exclusividad léxica.

Bloque	Nuevo lema	M / B	| N - M / B |
1	M: 3 427	3 427	0
2	1 646	1 714	68
3	1 263	1 142	121
4	1 080	857	223
5	759	685	74
6	775	571	204
7	552	490	62
8	523	428	95
9	560	381	179
10	401	343	58
11	488	312	176

12	335	286	49
13	347	264	83
14	335	245	90
15	263	228	35
16	292	214	78
17	268	202	66
18	268	190	78
19	195	180	15
20	199	171	28
Total	13 976	12 329	1 782

Tabla 2.4.a. Bloque y Nuevo lema en *Fortunata y Jacinta*

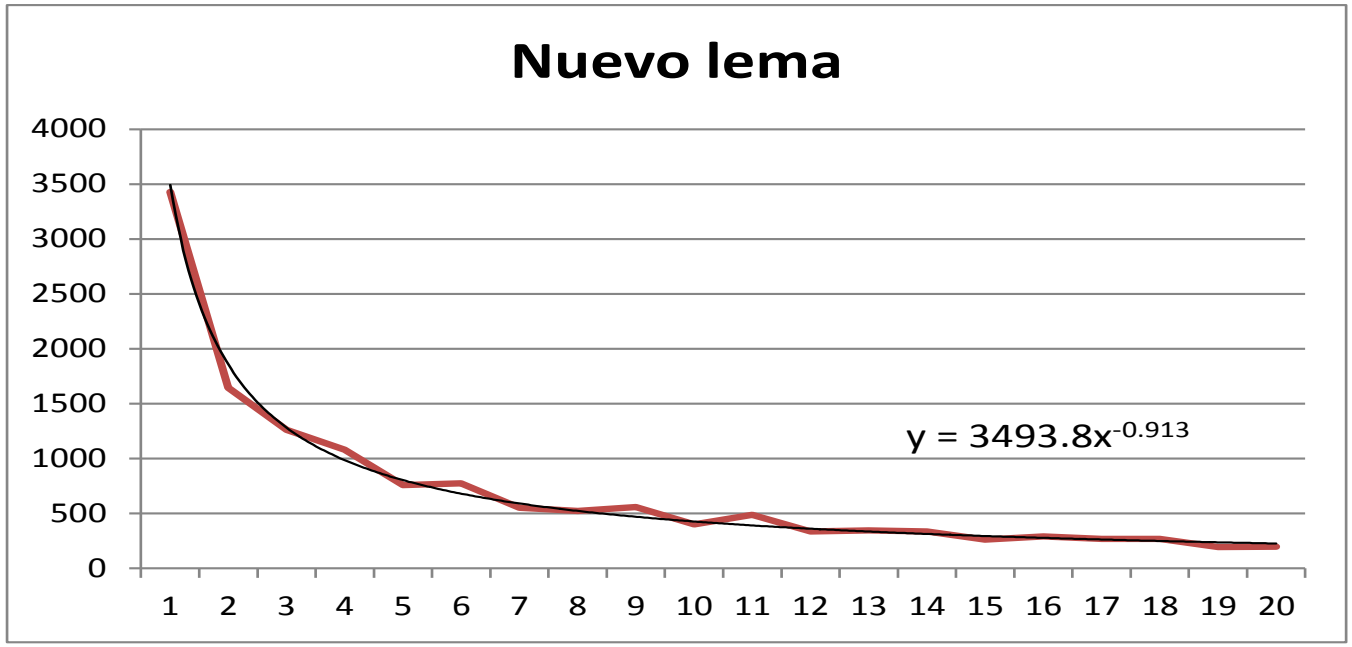


Fig. 2.4.a. Bloque y Nuevo lema en *Fortunata y Jacinta*

$y = 3493 * x^{-0.913} \rightarrow y = 3493 / x^{0.913}$

$y = 3493 / x \rightarrow x * y = 3493$

¿A qué se debe esta relación formulada?

Bloque	Nuevo lema	N.acum.
1	3 427	3 427
2	1 646	5 073
3	1 263	6 336
4	1 080	7 416
5	759	8 175
6	775	8 950
7	552	9 502
8	523	10 025
9	560	10 585
10	401	10 986
11	488	11 474

12	335	11 809
13	347	12 156
14	335	12 491
15	263	12 754
16	292	13 046
17	268	13 314
18	268	13 582
19	195	13 777
20	199	13 976

Tabla 2.4.b. Frecuencia y frecuencia acumulada de nuevos lemas en *Fortunata y Jacinta*

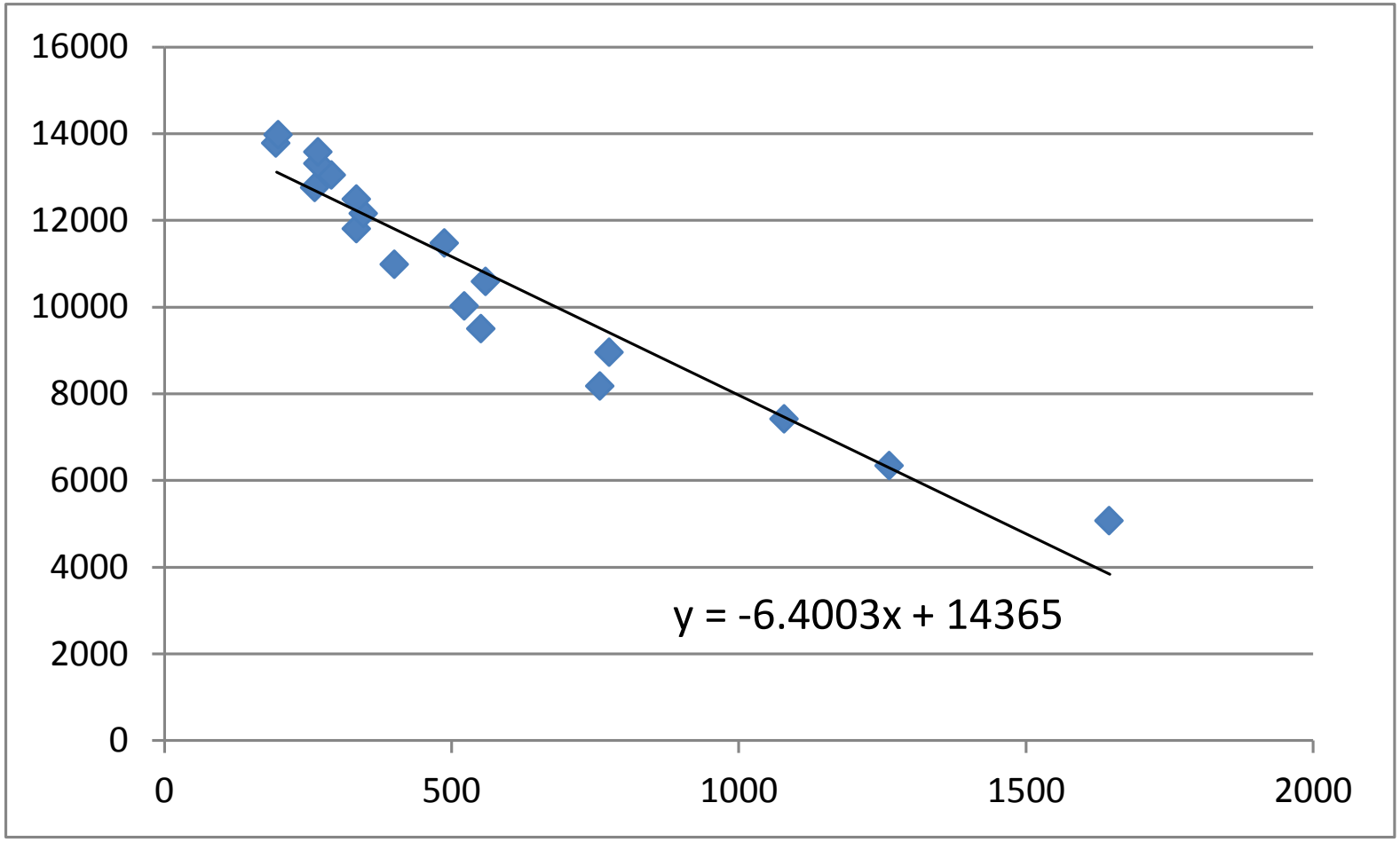


Fig. 2.4.b. Frecuencia y frecuencia acumulada de nuevos lemas en *Fortunata y Jacinta*

Bloque	a	abadesa	abajo	abalanzar	abalorio	abandonado	abandonar	abandono	abanicar	abanico	abarcar	abatido	abatimiento	abatir
1	571		2							3	1			
2	523	1	7				3	3						
3	591		3			1	1	1			1			
4	629		5	1			3							1
5	584		7				6						1	
6	591		4			1	3							1
7	594		1	1		1	2						1	2

8	598														1
9	584	6			1	1	1			1					
10	662	5	1	1		3									
11	659	3	1			6									
12	500	2				4					1				
13	548	5													1
14	645	5				1									1
15	573	5													
16	618	2					1		1			2			
17	565	1				2			4			2			
18	605	3				2									1
19	677	6				5		1	1						

20	680		4				5							
Total	11997	1	76	4	1	4	47	6	1	9	3	1	6	8

Tabla 2.4.c. Distribución de frecuencias de lemas en *Fortunata y Jacinta*

Bloque	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
abadesa		1																			1
abalorio										1											1
abanicar																			1		1
abatido												1									1
abdicar											1										1
abeja			1																		1
(…)	..	..	..																		1
Total	426	300	325	325	262	292	235	243	268	211	291	203	243	234	183	215	222	223	169	189	

Tabla 2.4.d. Distribución de frecuencias de lemas hápax en *Fortunata y Jacinta*

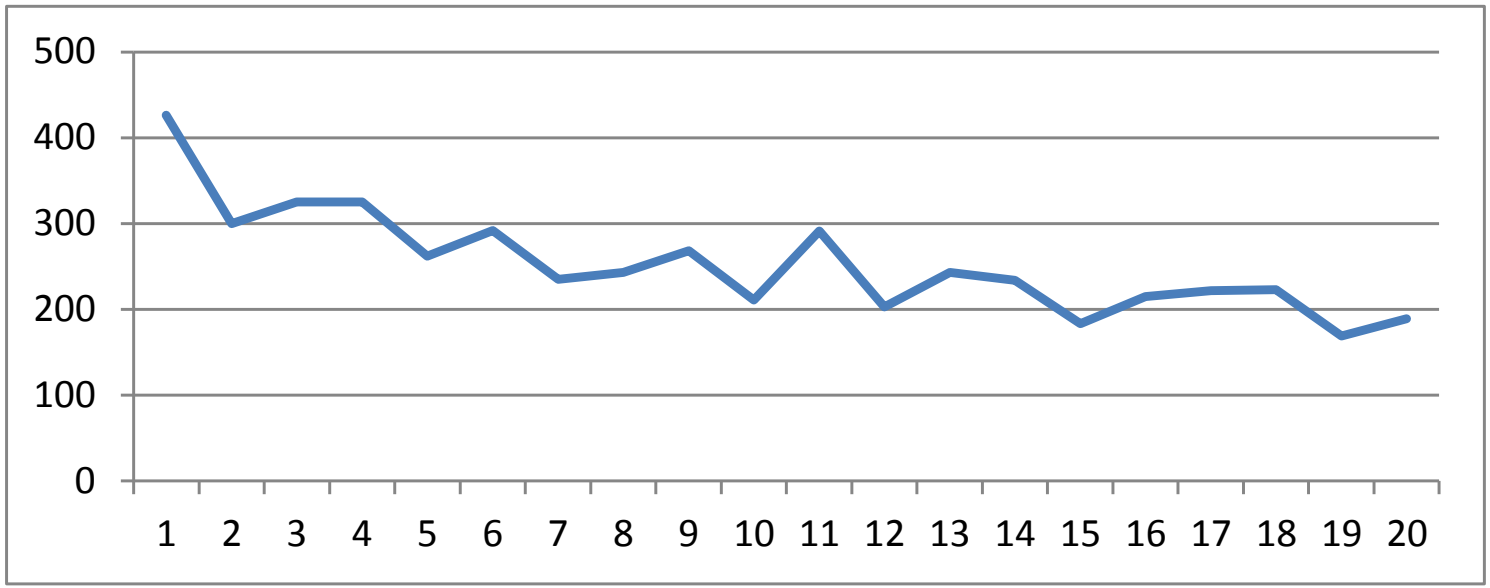


Fig. 2.4.c. Distribución de frecuencias totales de lemas hápax en *Fortunata y Jacinta*

Primeros hápax en orden alfabético en el bloque-20:

aberración, acerado, aderezado, afluir, alborotadamente, angélica, apalabrar, aparejar, apencar, artificioso, asquerosidad, asturiano, atisbador, batallita, bondadosamente, bozal, camposanto, canasto, cebollino, ciclón, colorete, compañerismo, confitar, cuajo, etc.

3. Conclusión

Total (de las formas o lemas contadas dentro del texto)
Tipo (números de formas o lemas diferentes)
Máximo (frecuencia máxima, la del artículo *el*)
Hápax (número de las formas o lemas de frecuencia única)
→
Ratio de Tipo / Total (ing. *Type / Token Ratio*: TTR)
Índice de Tipo (IT)
Índice de Hápax (IH).

Referencia

Baayen, R. Harald. (2001): *Word Frequency Distributions*. Dordrecht. Kluwer Academic Publishers.

Baker, Paul / Hardie, Andrew / McEnery, Tony. (2006): *A Glossary of Corpus Linguistics.* Edinburgh University Press.

Cresti, Emanuela / Moneglia Massimo (eds.) (2005). *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages.* Paris. John Benjamins

Miller, George A. (1951): *Language and Communication.* New York. McGraw-Hill. (traducción española por Eduardo

Goigorsky y Silvia Delpy, 1979) *Lenguaje y comunicación*. Buenos Aires. Amorrortu editores.

Moreno, Antonio. & Guirao, José María. (2006): "Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation", in *Spoken Language Corpus and Linguistic Informatics,* John Benjamins.

Moreno et al. (2005): "The Spanish Corpus", in Cresti & Moneglia (eds)

Real Academia Española. (2014): *Diccionario de la lengua española*. 23.ª ed. Madrid: Espasa. http://dle.rae.es/?w=diccionario [26/mayo/2018]

Stubbs, Michael. (2002): *Words and Phrases. Corpus Studies of Lexical Semantics.* Oxford. Blackwells.

Ueda, Hiroto / Moreno Sandoval, Antonio. (2018): *Sistema LYNEAL «Letras y Números en Análisis Lingüísticos»* https://lecture.ecc.u-tokyo.ac.jp/~cuedályneal/ [24/ abril/2018] http://shimoda.lllf.uam.es/uedályneal/ [24/ abril/2018]

Zipf, George Kingsley. (1936): *The Psycho-bilology of language.* New York. George Routladge & Sons.

____. (1949): *Human Behavior and the Principle of Least Effort*. Cambridgé Massachusetts: Addison-Wesley.