Algunhas calas na riqueza léxica da lingua literaria: aproximación cuantitativa

XOSÉ L. REGUEIRA
Instituto da Lingua Galega

De cando en vez atopámonos con opinións ou xuízos sobre o vocabulario utilizado nas obras literarias, tanto sobre a súa cantidade coma sobre a variedade ou o tipo de palabras empregadas. Sen embargo, a penas contamos con estudios nin con datos seguros sobre o acervo léxico dos autores literarios galegos, fóra dalgúns glosarios, poucas veces exhaustivos.

Este breve traballo constitúe unha aproximación cuantitativa ó léxico dalgunhas obras da narrativa galega das últimas décadas. Naturalmente, a riqueza léxica dunha obra non se mide soamente na cantidade de palabras diferentes que utiliza, senón que tamén habería que ver se o léxico usado corresponde con aquel máis frecuente no uso cotián da lingua ou se polo contrario se manexa léxico de uso limitado (arcaísmos, dialectalismos, vulgarismos, léxico específico de certas actividades ou esferas de coñecemento, etc.), e tamén comprobar a que campos semánticos corresponde o núcleo léxico fundamental de cada obra, para logo poñer esa información en relación coa intención estética do escritor[1]. Polo de pronto imos estudia-lo léxico de catro obras narrativas unicamente no seu aspecto cuantitativo. Cando chegue a termo a base de datos lexicográfica que se está a realizar no Instituto da Lingua Galega baixo a dirección do Prof. Antón Santamarina, poderá abordarse con relativa facilidade o estudio tanto cuantitativo coma cualitativo dun elevado número de obras da literatura galega.

Os textos narrativos analizados foron[2]:

[1] Cfr. un pequeno ensaio nesa dirección en Regueira (1987) e Sánchez Palomino (1987).

[2] Quero agradecerlle ó equipo de investigadores do proxecto *Lexicografía* que se desenvolve no Instituto da Lingua Galega, en especial ó seu director Antón Santamarina e mais a Margarita Neira, a súa xenerosidade ó permitirme utiliza-los textos e as bases de datos lematizadas que serviron de base para o cómputo de case todas estas obras. Tamén lle agradezo a Arturo Reguera a súa inestimable axuda na solución dos problemas informáticos que se foron presentando.

Cunqueiro, A.: *Merlín e familia (i outras historias)*. Vigo: Galaxia, 1968.

Blanco Amor, E.: *Xente ao lonxe*. Vigo: Galaxia, 1988.

Cabana, D. X.: *Galván en Saor*. Vigo: Xerais, 1989.

Méndez Ferrín, X. L.: *Arraianos*. Vigo: Xerais, 1991.

Efectuouse tamén o cómputo do vocabulario de obras doutro carácter para que servisen como referencias de contraste: a peza de D. R. Castelao, *Os vellos non deben de namorarse* (Vigo: Galaxia, 1953), o libro de poemas de D. X. Cabana, *Patria do mar* (Vigo: Ir Indo, 1989) e finalmente unha colección de textos orais recollidos por min no inverno 1983-84 en varias parroquias dos concellos de Vilalba, Abadín e Xermade[3].

1. LEMATIZACIÓN

En primeiro lugar quixemos comproba-lo número de palabras diferentes utilizadas nas obras tomadas en consideración. Para iso, e dado que pretendemos un estudio cuantitativo do léxico, consideramos indispensable lematiza-las ocorrencias léxicas, é dicir, reducir a lemas as formas empregadas no texto, prescindindo da flexión nominal de xénero e número e da flexión verbal, para despois ordenalas.

Non tomamos en consideración as opinións dalgúns investigadores que critican a reducción das formas flexionadas a lemas aducindo que a información de tipo gramatical e mesmo fonético que se perde é moi importante (Cfr. Irizarry 1990: 269). É evidente que, para a finalidade proposta, a dispersión de formas producida polas flexións nominal e verbal, así como a importante cantidade de homógrafos indiferenciados, produciría tal distorsión que os datos resultarían de moi pouca utilidade (V. a discusión en Muller 1984).

A lematización implica tamén outra tarefa máis delicada: a desambiguación de homógrafos. Por contra, os casos de polisemia mantivéronse coma unha entrada única. Tamén se tomaron como entradas diferenciadas as contraccións consolidadas (*á, do, nalgún*, etc.).

É sabido que non adoita haber acordo entre os chamados con certo humor "contadores de palabras" (fr. *compteux de mots*) á hora de decidir onde

[3] Nas parroquias de Alba, Belesar, Boizán, Carballido, Codesido, Corbelle, Distriz e Goiriz, do concello de Vilalba; en Aldixe e Castromaior, de Abadín; en Burgás, Cabreiros, Candamil e Cazás, no concello de Xermade.

se debe traza-lo límite entre as formas que han de ser consideradas palabras independentes e outras que se tratan como parte de sintagmas fixados de contido semántico unitario, ou á hora de traza-los límites da desambiguación (Muller 1984: V-VI). Ó encara-lo traballo práctico aparece outra serie de problemas: ¿deben considerarse palabras diferentes *ser* utilizado como verbo e *ser* empregado como substantivo? ¿E *alto* como adxectivo e como adverbio? ¿Cales son os criterios que deben decidir en que casos un participio ten xa existencia independente como adxectivo? ¿Deben ser considerados entradas independentes tódolos derivados ou deben excluírse os diminutivos e aumentativos? De tódolos xeitos, son problemas de importancia relativa, xa que consideramos que, calquera que sexa a decisión que se tome, será válida sempre que se aplique de maneira sistemática a tódalas obras que se van comparar. Ademais, as variacións cuantitativas que produciría a adopción dunha postura ou outra serían pouco relevantes. O criterio que seguimos é distinguir como lemas diferentes todas aquelas formas que figuran como entradas independentes no diccionario (Cfr. Muller 1968: 253-259).

2. PROCEDEMENTO E CUESTIÓNS DE MÉTODO

Tralo reconto das ocorrencias (N = número total de empregos de palabras) e despois de resoltos os problemas de lematización, obte-las cifras de vocabulario (V = número de palabras utilizadas polo menos unha vez no texto) non presenta grandes dificultades. Nas obras estudiadas obtémo-los seguintes resultados:

	N	V	Porcentaxe V/N
Merlín	25741	2945	11.44 %
Arraianos	34822	5222	14.99 %
Galván	45177	4104	9.08 %
Xente	74089	5886	7.94 %
Vellos	10291	1463	14.21 %
Patria	11030	2592	23.49 %
Gravacións	58301	3235	5.54 %

Estas cifras mostran que o 11 % do texto do *Merlín* é vocabulario novo, utilizado por primeira vez, isto é, case o 89 % restante son ocorrencias

repetidas. Estes resultados indícannos que a porcentaxe de V é maior, p.e., nos *Vellos* ca en *Xente ó lonxe*. Pero, tomados así, estes datos poderían levarnos a conclusións erróneas, xa que estas porcentaxes só poden ser comparadas entre dúas obras de extensión (N) igual ou moi semellante.

Tomando un exemplo suposto, se un texto de 25000 ocorrencias léxicas (N) contén un vocabulario (V) de 3000 entradas, non podemos esperar que outro texto do dobre de extensión (50000 palabras), e dunha riqueza léxica semellante, teña o un vocabulario de 6000 voces. Como é evidente, a medida que avanzamos no texto aumentan as posibilidades de que unha palabra se repita, e polo tanto diminúen as posibilidades de que aparezan palabras novas. Ó inicio dun texto mantense V = N ata que aparece a primeira repetición, que provoca un atraso dunha unidade en V respecto de N. A medida que se producen repeticións V atrásase máis respecto de N, e cada vez encontraremos máis ocorrencias repetidas e os intervalos de aparición de voces novas serán maiores (Cfr. Muller 1968: 268).

É claro que a medida que crece N tamén vai aumentando V. "Pero, ¿según qué ley? No se sabe todavía con precisión" (Muller 1968: 267). Canto máis se incrementa N máis lento é o crecemento de V, i.e., canto máis avanzamos no texto menos palabras novas aparecen. "Pero V no cesa de crecer, pues ningún texto agota el léxico de su autor" (Muller 1968: 268), polo que a liña de V tenderá a facerse paralela co eixe de abscisas pero sen chegar a selo totalmente, dunha maneira semellante ó gráfico 2, máis abaixo. Ou, presentado doutra maneira, se contámo-lo número de palabras que se usan por primeira vez en intervalos regulares nun texto, e denominamos V_1 ó vocabulario novo do primeiro tramo, V_2 ó do segundo, etc., podemos prever que $V_1 > V_2 > V_3 ... > V_n$.

É posible aproximarse, a través do cálculo de probabilidades, ó crecemento esperado do vocabulario nun texto dado, unha vez coñecidos os valores de N e V para todo o texto, así como o número de palabras con frecuencia 1, 2, 3... n. Este cálculo, de execución traballosa, é exemplificado por Muller (1968: 310-322), xa que a contaxe directa do incremento de V en relación co incremento de N é considerada inviable (p. 310): "Es necesario [...] figurarse un lector dotado de una atención y de una memoria sobrehumanas, que fuese capaz, a todo lo largo del texto, de contar a la vez las palabras (que constituyen N) y los vocablos (que forman V), notando cada vocablo nuevo que viene a incrementar V en una unidad". Hoxe os medios informáticos fan posible esa tarefa.

Para obter esa información creamos unha base de datos na que cada palabra se asocia co seu número de orde de aparición no texto. Trala lemati-

zación, creamos un índice coa primeira ocorrencia de cada lema. Isto permítenos coñecer con exactitude en que momento se incorpora ó texto cada unha das palabras do vocabulario, e polo tanto saber cantas palabras se utilizan por primeira vez en cada un dos tramos que queiramos establecer nun texto[4]. Tendo en conta as consideracións expostas máis arriba, é de esperar que desta contaxe resulten curvas de aparición de V semellantes a esta (no eixo de abscisas os tramos de texto de mil en mil palabras e no de ordenadas o número de vocábulos de primeira aparición):

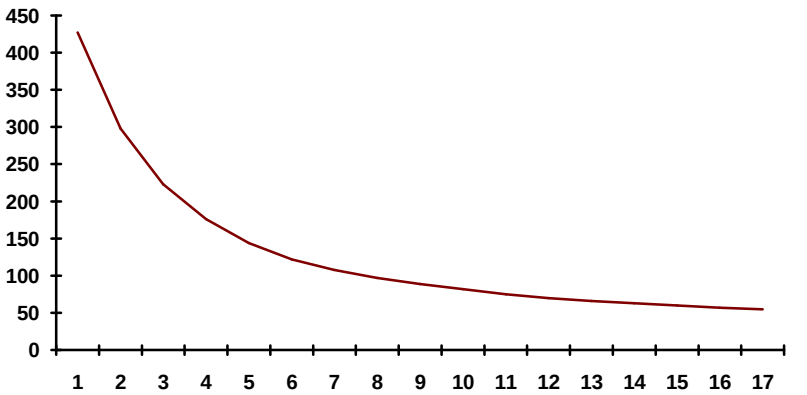

gráfico 1

Unha vez coñecidos estes datos, non temos máis ca face-la suma acumulada dos valores de cada tramo para obtérmo-la curva de crecemento de V para a obra considerada, que para os datos presentados no gráfico 1 será esta:

[4] Traballamos con tramos de 1000 e de 5000 palabras. Para os gráficos prescindimos das fraccións menores; p.e., se *Galván* ten unha extensión de 45177 palabras, tomamos en consideración as 45000 primeiras ocorrencias e prescindimos das 177 ocorrencias restantes.

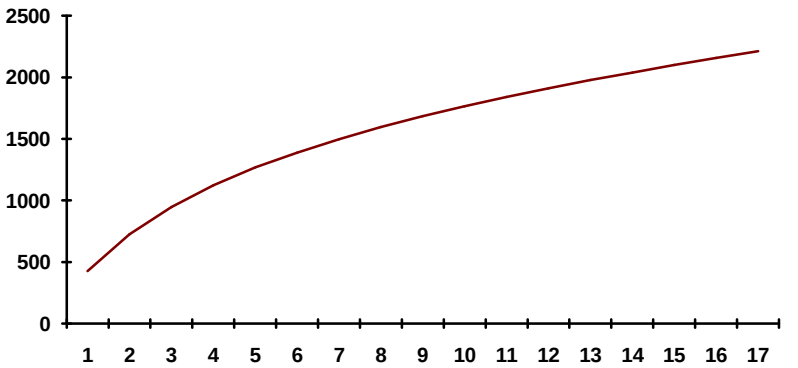

gráfico 2

As curvas de crecemento de V (gráfico 2) e de incorporación de léxico (gráfico 1) obtidas en distintas obras poden ser comparadas independentemente da súa extensión, xa que nos permiten saber cal é o valor de V para cada momento de N; é dicir, podemos saber con precisión cantas voces novas aparecen entre dúas ocorrencias calquera (p.e. entre a palabra 20000 e a palabra 20500), e tamén sabemos cantos vocábulos diferentes aparecen p.e. nas 10000 primeiras palabras dunha obra calquera que teña como mínimo esa extensión.

Esta contaxe do vocabulario de cada tramo permite ademais aprecia-las alteracións na progresión de V (i.e., aqueles treitos en que o volume de incorporacións léxicas é menor ou maior do esperado) que, debido a variacións estilísticas ou temáticas, se produzan ó longo do texto. Cabe esperar que a introducción de novos temas, de novas situacións que obriguen a describir, levará consigo un incremento acusado do vocabulario, mentres que a reiteración de situacións e temas tratados fará que o incremento do vocabulario sexa máis lento. Da mesma maneira, as variacións no estilo de lingua (coloquial, popular, culta, específica dun ámbito social concreto, etc.) afectarán á cantidade de vocábulos empregada. Por iso as curvas que resultan da contaxe levada a cabo presentan un aspecto moito máis irregular cás curvas ideais presentadas nos gráficos 1 e 2.

3. CRECEMENTO DO VOCABULARIO NAS OBRAS ESTUDIADAS

Seguindo o procedemento indicado (cfr. gráfico 2), trazámo-las curvas de crecemento de V das catro obras narrativas do noso corpus:

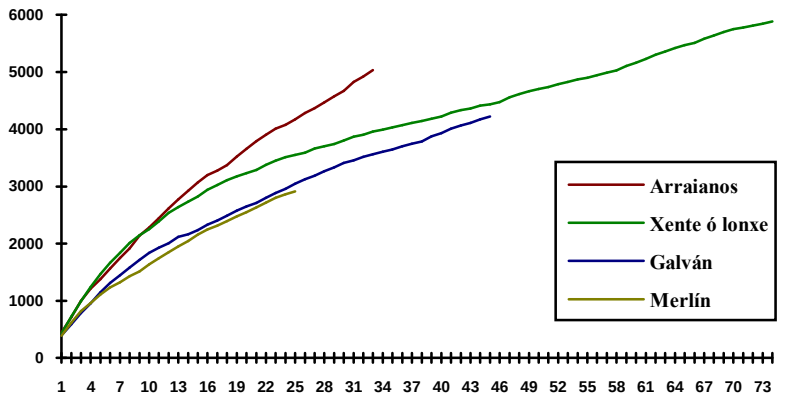

Gráfico 3
V acumulado das catro obras narrativas

Este gráfico permítenos facer unha serie de observacións interesantes. Se tomamos en conta que a orde das curvas corresponde coa da lenda, vese ben que a obra máis rica en vocabulario é *Arraianos*, seguida de *Xente ó lonxe* e, máis distanciadas, *Galván* e *Merlín*. As dúas primeiras mantéñense moi próximas nos primeiros miles de palabras, e a novela de Blanco Amor supera á de Ferrín en léxico acumulado entre as 4000 e as 9000 palabras; pero de seguido en *Xente* prodúcese unha caída apreciable no ritmo de crecemento do vocabulario, en tanto que en *Arraianos* mantén un ritmo de crecemento de V alto e constante ata o final da obra, de maneira que ás 33000 palabras a obra de Ferrín presenta unha vantaxe de 1074 voces (*Arraianos* 5032, *Xente* 3958). Esa caída de *Xente ó lonxe*, prolongada ó longo de bastantes miles de palabras, fai que case sexa alcanzada por *Galván en Saor*; despois de levar unha vantaxe de 622 voces á altura da palabra 17000 (*Xente* 3026, *Galván* 2404), a distancia vaise reducindo ata quedar só en 212 voces cando remata a obra de Cabana, ás 45000 palabras (*Xente* 4435, *Galván* 4223). *Merlín e familia* mantén un nivel de incremento léxico moi semellante ó de *Galván*, e

tras superalo levemente nos primeiros miles de palabras, atrásase un pouco e logo segue unha traxectoria paralela á novela de Cabana, para acabar ás 25000 palabras con 136 voces menos (*Galván* 3046, *Merlín* 2910).

Para tomar unha referencia que nos dese unha idea da riqueza léxica dos textos utilizados respecto da lingua oral, fixémo-lo mesmo cómputo con textos de gravacións cunha extensión de 58000 palabras. Como é evidente, no discurso oral prodúcense repeticións moi frecuentes, especialmente de certas palabras baleiras de significado (*bueno, pois,* etc.), así como dalgunhas palabras gramaticais, e téndese ó emprego reiterado dun número reducido de palabras léxicas de significación moi extensa (*haber, facer, cousa,* etc.). Por iso era previsible unha distancia considerable na riqueza relativa de vocabulario novo, que se ve confirmada polos datos obtidos. Tamén contámo-lo vocabulario dunha peza teatral, *Os vellos non deben de namorarse,* que polo seu carácter dialogado e polo ambiente no que se desenvolve cabe esperar que se encontre máis próxima á fala oral cás obras narrativas. Isto tamén se reflicte neste gráfico, que mostra os *Vellos*, de só 10000 palabras de extensión, nunha posición intermedia entre a curva de *Merlín* e a das gravacións (con todo o seu afastamento dos valores da lingua oral é unha proba da elaboración literaria deste texto):

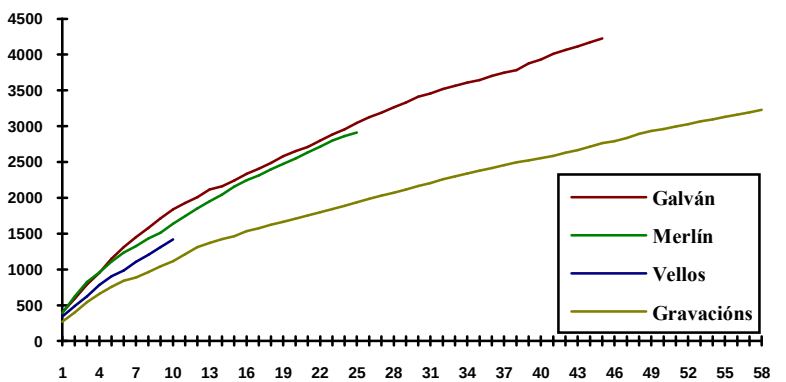

Gráfico 4

V acumulado de Galván, Merlín, Vellos e gravacións

Por último quixemos facer unha cala na lingua utilizada na poesía, para tomar outro punto de referencia. Parece razoable prever que a riqueza léxica

da poesía será considerablemente maior cá da narrativa. Para axustar mellor o contraste, escollemos *Patria do mar* de Darío Xohán Cabana. Este poemario é obra do mesmo autor dunha das novelas analizadas, *Galván en Saor*, foi publicado no mesmo ano e, ademais, unha boa parte dos poemas trata temas comúns á novela; desta maneira podemos pensar que as diferencias encontradas se deben fundamentalmente ó diferente carácter dos textos poético e narrativo. O gráfico compara o incremento acumulado do vocabulario de *Patria* coas 11000 primeiras palabras de *Galván* e da obra narrativa máis rica, *Arraianos*:

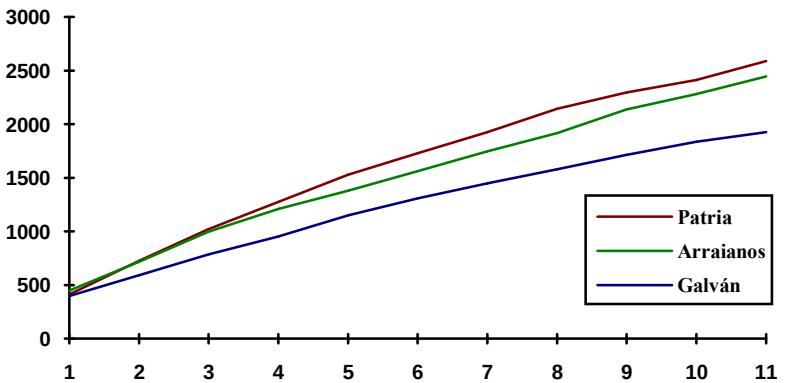

Gráfico 5
V acumulado de Patria, Arraianos, Galván (11000 primeiras palabras)

Neste gráfico apréciase como a liña de *Patria do mar* se afasta da seguida por *Galván en Saor* desde o principio e como a separación entre as dúas se vai facendo cada vez maior. *Arraianos* supera a *Patria* no primeiro milleiro de palabras, para logo perder posicións, aínda que na parte final as liñas de ambas obras se manteñen paralelas. Os valores de V para as 11000 primeiras palabras son de 2589 voces en *Patria*, 2445 en *Arraianos* e 1928 en *Galván en Saor*.

4. VARIACIÓNS ESTILÍSTICAS NO CRECEMENTO DE V

Como indicamos máis arriba, as curvas de incorporación de léxico novo fornécennos información sobre variacións estilísticas ou temáticas dentro dunha obra. Se en lugar de considera-lo vocabulario acumulado tomámo-los valores de V para cada treito de mil palabras (cfr. gráfico 1), podemos observa-las variacións que se producen no crecemento de V en *Galván en Saor*:

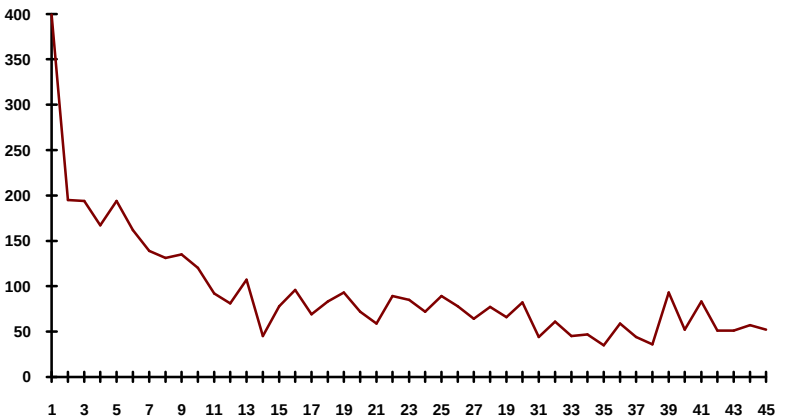

Gráfico 6
Galván en Saor

Neste gráfico vemos numerosas variacións de curto alcance, que se presentan coa forma de picos ou de depresións máis ou menos bruscas. A nosa base de datos permítenos observa-las variacións con detalle, e así comprobamos que a depresión relativa tan profunda (V = 45) que se encontra no milleiro 14 corresponde a un diálogo no castelo de Mirez entre Galván, Silvania e Vedromil (pp. 42-44). A parte máis pobre en vocabulario novo é a que vai da palabra 30000 á 38000, dedicada a narra-la partida cara o sur dos protagonistas, o seu encontro co ermitán, a estadía na pousada e outros episodios que ocupan os capítulos 10 e 11, nos que abunda o diálogo. Das 38000 ás 39000 ocorrencias prodúcese un pico de 93 voces utilizadas por primeira vez, debido ó elevado número de vocábulos empregados en describi-los elementos da paisaxe no capítulo 12. Tras un declive relativo de 52 voces novas entre o final deste capítulo e o comezo do seguinte ata a entrada dos protago-

nistas na caverna da montaña, encontramos outro cume de 83 vocábulos novos entre as ocorrencias 40000 e 41000, nun anaco rico en descricións que inclúe a loita de Galván coa serpe.

Antes viamos (gráfico 3) que a curva de V acumulado en *Xente ó lonxe* presentaba un aplanamento relativo entre as 30000 e as 45000 palabras, aproximadamente, para recuperar posteriormente valores máis altos, sobre todo a partir da palabra 59000. Se considerámo-lo vocabulario novo (V) por cada 1000 palabras, obtemos:

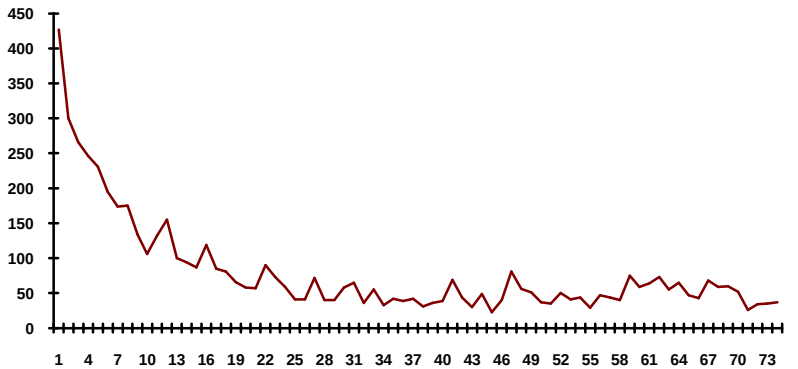

Gráfico 7

Xente ó lonxe

En xeral, vese que se rexistran valores baixos para V entre os milleiros 25 e 58, que corresponden a partes nas que predomina o diálogo e o tratamento de temas xa mencionados previamente. O pico relativamente prominente que se rexistra entre as palabras 26000 e 27000 (V = 72) débese á abundancia de léxico relativo á comida e de termos empregados para a descrición de Evanxelina no comezo do capítulo 7 da primeira parte. Outra elevación importante (V = 81), no milleiro 47, corresponde ó final do capítulo 3 da segunda parte, no que se presenta unha discusión política rica en cultismos. A partir da palabra 58000 prodúcese unha elevación xeneralizada dos valores de V, nunha parte en que predomina a narración sobre o diálogo, e logo volven a diminuír a partir da palabra 70000, xa no final da obra.

Ese cambio que afecta a períodos longos pódese apreciar mellor se tomamos treitos máis extensos (5000 ocorrencias), xa que as variacións de curta extensión resultan así menos visibles:

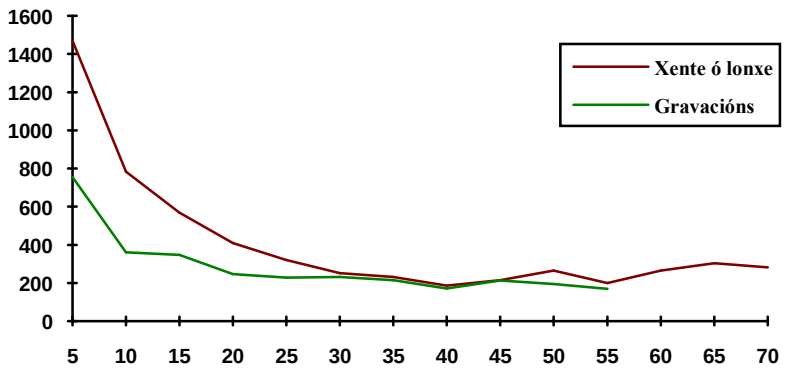

Gráfico 8
Xente ó lonxe / Gravacións

Vese aquí como a curva de incremento de vocabulario de *Xente ó lonxe* cae ata poñerse ó mesmo nivel da obtida nas gravacións orais no treito comprendido entre os 30 e os 45 milleiros de palabras, para logo remontar e presentar valores algo máis altos nas 25000 palabras seguintes.

Se no gráfico 3 viamos como, gracias ó escaso crecemento de V nese treito, a curva de vocabulario acumulado de *Xente ó lonxe* se afastaba da de *Arraianos* e se achegaba á de *Galván en Saor*, podemos apreciar agora estas variacións con maior detalle presentando o crecemento de V en treitos de 5000 palabras para as tres obras:

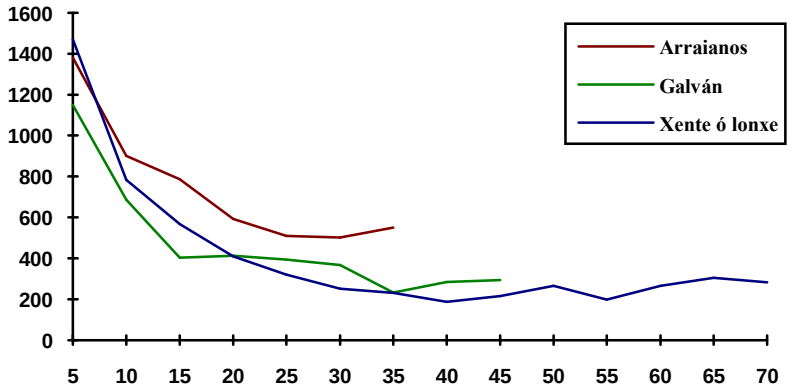

Gráfico 9
Arraianos / Galván / Xente ó lonxe

Obsérvase ben que *Xente ó lonxe* é a obra que obtén un número máis alto de vocábulos nas primeiras 5000 palabras para ser superada de contado por *Arraianos*, que vai presentar valores claramente superiores en tódolos tramos restantes; a curva de V da novela de Blanco Amor segue caendo ata verse superada a partir das 20000 palabras por *Galván*. Como vimos anteriormente, isto produce un acercamento entre as curvas de vocabulario acumulado de ambas novelas, pero sen que a de Cabana chegue nunca a superar *Xente ó lonxe* (gráfico 3).

5. CONSIDERACIÓNS FINAIS

Nesta exposición comprobamos que o procedemento de contaxe de palabras utilizado permite comparar con gran precisión o vocabulario utilizado en obras literarias, independentemente da súa extensión. Os datos presentados mostran con claridade o diferente nivel de utilización dos recursos léxicos da lingua por distintos autores, e neste sentido corrobora diferencias léxicas que o lector atento á lingua pode percibir de maneira intuitiva. Un achegamento máis completo e máis enriquecedor á utilización do léxico require, como diciamos na introducción, unha análise cualitativa do vocabulario empregado.

Outra das posibilidades deste procedemento é a detección de diferencias marcadas na utilización do vocabulario ó longo dun texto, tanto nunha obra tomada por si soa, como comprobamos no comentario de *Galván en Saor* e de *Xente ó lonxe*, coma establecendo comparacións entre o desenvolvemento de varias obras diferentes. Isto permitiría confronta-lo crecemento do vocabulario en textos de distinto carácter, en varias obras dun mesmo autor ou, como fixemos aquí, en obras de autores diferentes.

BIBLIOGRAFÍA CITADA:

Irizarry, E. (1990): "Stylistic analysis of a corpus of twentieth-century Spanish narrative", *Computers and the Humanities* 24, 4, pp. 265-274.

Muller, C. (1968): *Initiation à la statistique linguistique*. Paris: Larousse. Cito pola trad. esp.: *Estadística lingüística*. Madrid: Gredos, 1973.

Muller, C. (1984): "De la lemmatisation", prefacio a P. Lafon (1984): *Dépuillements et statistiques en lexicométrie*. Génève / Paris: Slatkine / Champion, pp. I-XII.

Regueira, X. L. (1987): "O galego de *Nimbos*", in *Homenaxe a X. M. Díaz Castro*. Guitiriz: Xermolos, pp. 85-102.

Sánchez Palomino, M. D. (1987): "Aproximación a unha análise estilístico-formal de *Nimbos*", in *Homenaxe a X. M. Díaz Castro*. Guitiriz: Xermolos, pp. 114-149.