

ÍNDEX DE DESIGUALTAT LINGÜÍSTICA DE LES REGIONS SOCIOLINGÜÍSTIQUES VALENCIANES

21 de febrer del 2024

Dia Internacional de la Llengua Materna (UNESCO)

Síntesi

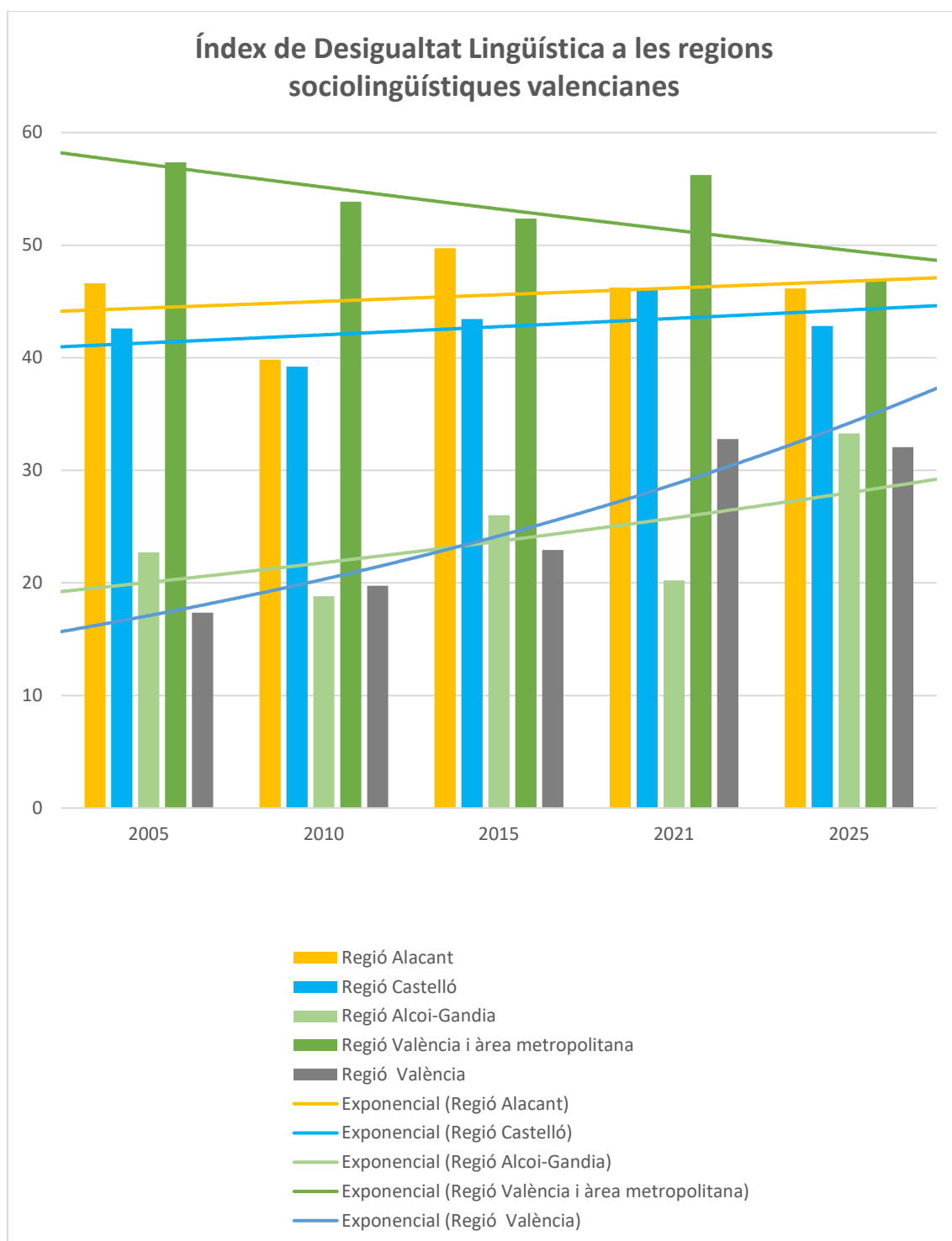
La Fundació Nexa adverteix que **la desigualtat lingüística referida al valencià creixerà en la major part de regions sociolingüístiques valencianes**, especialment a les regions que ara presenten una desigualtat menor.

Els resultats de l'Índex de Desigualtat Lingüística, referit al valencià, s'arreglen en la taula següent:

	2005	2010	2015	2021	2025*	TENDÈNCIA
Regió Alacant	46,61	39,83	49,72	46,23	46,17	creixent
Regió Castelló	42,59	39,21	43,44	46,01	42,82	creixent
Regió d'Alcoi-Gandia	22,71	18,81	26,02	20,22	33,27	creixent
Regió de València i àrea metropolitana	57,34	53,86	52,35	56,23	46,91	decreixent
Regió de València	17,36	19,75	22,92	32,78	32,05	creixent

*Estimació amb IA.

Aquestes dades es poden representar al gràfic següent:



L'Índex de Desigualtat Lingüística (IDL) que presentem hui és una adaptació de l'Índex de Gini, que és l'estàndard per a avaluar la desigualtat socioeconòmica, i mesura l'evolució de la desigualtat entre les competències i usos públics del valencià. Aquest indicador ens permet fer-hi comparacions entre les situacions en les diverses regions lingüístiques, així com avaluar quina és la seua evolució passada i fer previsions sobre els possibles futurs que ens esperen.

Les regions sociolingüístiques amb més desigualtat lingüística són València i la seua àrea metropolitana, seguides de la regió d'Alacant valencianoparlant. L'estudi de les sèries de l'Enquestes de Coneixement i Ús del Valencià, així com una prognosi dels resultats del 2025, realitzada amb Intel·ligència Artificial, permet concloure que hi ha una notable tendència creixent de la desigualtat lingüística en les regions sociolingüístiques on ara és menor, a la regió d'Alacant i, sobretot, a les zones valencianoparlants de les regions de Castelló i València i a la regió Alcoi-Gandia. És a dir, la major part de les regions sociolingüístiques presenten tendències a l'augment de la desigualtat fins i tot més enllà del 2025. Per això, *si no es prenen mesures de política lingüística efectiva, assistirem a un increment general de la desigualtat lingüística.*

Susanna Pardines, responsable de Nexa, ha declarat: «La nostra Fundació està molt preocupada per si les noves autoritats no continuen amb els estudis sobre la situació del valencià. S'ha aturat la convocatòria d'estudis d'enguany i contemplem amb molta inquietud que es puguin interrompre les sèries d'enquestes. El coneixement de les dades de competències, usos i expectatives és imprescindible per a la revitalització del valencià».

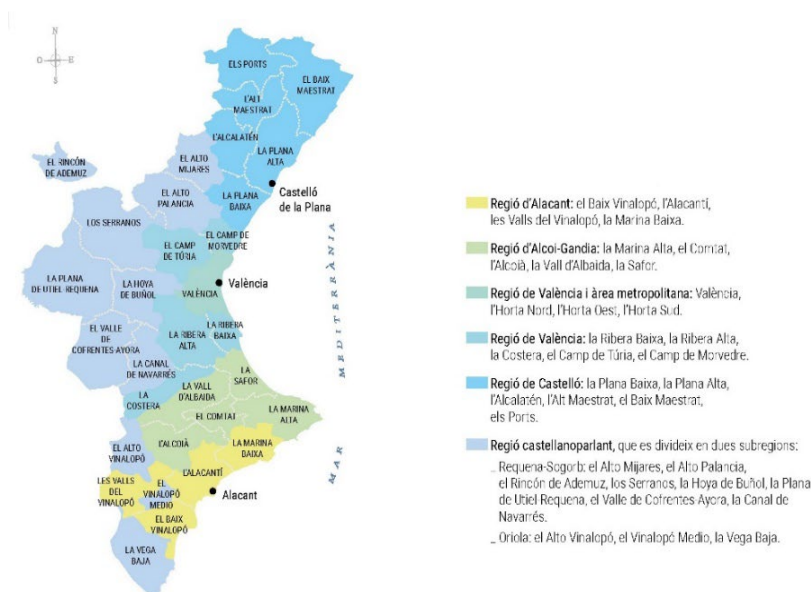
D'altra banda, Marc Hernández, responsable d'Ethos Insight, ha afirmat que el càlcul de l'IDL s'ha realitzat amb les dades depurades de les Enquestes de Coneixement i Ús del Valencià de 2005, 2010, 2015 i 2021. Per a les dades del 2025 s'ha emprat la metodologia *Random Forest* d'Intel·ligència Artificial. «Tenim experiència en estudis sobre el valencià i el nostre objectiu és continuar amb aquestes recerques i aplicar-ne les metodologies més avançades».

La realitat sociolingüística del valencià està caracteritzada per un creixement de les competències en valencià de la població, amb una caiguda dels seus usos socials. Una desigualtat que no es dona amb les competències i usos del castellà.

Aquesta desigualtat en els processos d'evolució de les competències i els usos del valencià indica la presència de pressions socials i polítiques estructurals contra l'ús públic de valencià, malgrat ser-hi cada vegada més competents. Això provoca una situació d'idealització del valencià que distorsiona la percepció social de la situació de retrocés del valencià, i a la vegada genera una especulació amb les acreditacions de coneixements del valencià i una pujada falsa del seu valor. I aquests són uns dels principals problemes amb què s'hi troba la recuperació d'espais efectius d'ús social del valencià.

És per aquesta raó que fer un seguiment de quina és la situació de la desigualtat entre les competències i usos del valencià és molt rellevant per poder avaluar quina és l'evolució dels processos de manteniment i recuperació social del valencià. En aquest sentit, trobem fonamental poder disposar d'indicadors que assenyalen quins són els camins que recorrem respecte a la llengua pròpia de la societat valenciana.

La Fundació Nexa és una entitat sense ànim de lucre que promou recerques i estudis a favor de la cultura valenciana. Ethos Insight és una iniciativa d'anàlisi social, compromesa en projectes de valor ètic.



ANNEX METODOLÒGIC

- 1.- La mesura de la desigualtat
- 2.- Adaptació de l'Índex de Gini en un Índex de Desigualtat Lingüística
- 3.- Les dades de les regions sociolingüístiques valencianes
- 4.- Ampliació i debat
- 5.- Pronòstic de dades amb Intel·ligència Artificial
- 6.- Bibliografia
- 7.- Llista de taules, gràfics, requadres i imatges

1.- La mesura de la desigualtat

La desigualtat social és un assumpte clàssic de la sociologia. Podríem dir, fins i tot, que aquesta disciplina es formulà a partir de la sospita que els ideals proclamats per la Revolució Francesa (llibertat, igualtat i fraternitat) no s'havien realitzat. Precisament perquè s'havia enunciat aquest ideal en l'autorepresentació del moviment revolucionari, la noció de desigualtat heretà un principi normatiu, que permet distingir-la de la mera descripció etnogràfica de les diferències socials. Aquesta càrrega normativa és la que està subjacent a nocions com ara *classe* (a diferència de l'*estrat* social), *gènere* (per contraposició a *sexe*) o *ètnia* (a diferència de l'acientífica *raça*).

Com és conegut, la desigualtat social es mesura per mitjà de dos procediments: la comparació de quantils i l'índex de Gini.

Un quantil és una part d'una mostra ordenada. Així, si la mostra es divideix per 5, parlem de quintils; per 10, decils, per 100, centils o percentils, etc. La comparació de quantils més habitual com a indicador de desigualtat es pot realitzar de dues maneres. La primera és amb la comparació de les mitjanes dels quintils. Així, per exemple, la Unió Europea mesura la desigualtat com la proporció (*share*) dels ingressos mitjans del quintil superior respecte de l'inferior. Així, la mitjana d'ingressos del quintil superior es representa com a *S80* (val a dir, la submostra més enllà del punt corresponent al 80% de la mostra ordenada si la representàrem en un eix) i la del quintil inferior com a *S20* (mitjana d'ingressos del primer quintil, els que estan a l'esquerra del 20%). D'aquesta manera, l'indicador de desigualtat es formula com: $S80/S20$ (EUROSTAT 2023).

Resulta trivial que és equivalent aquesta comparació si fem servir la mitjana d'ingressos o el conjunt dels ingressos de cada quantil. En el primer cas, la desigualtat δ serà:

$$\delta = \frac{S80}{S20}$$

Si per a una població de N individus, ens referim al total dels ingressos de cada quintil i anomenem, per exemple, $T80$ al total del quintil superior i $T20$ al total del quintil inferior, aleshores la proporció, que anomenarem δ' , serà:

$$\delta' = \frac{T80}{T20}$$

Si dividim el numerador i el denominador per la població de cada quintil, val a dir $\frac{N}{5}$, aleshores:

$$\delta' = \frac{\frac{T80}{\frac{N}{5}}}{\frac{T20}{\frac{N}{5}}} = \frac{S80}{S20}$$

Per tant, resulta trivial que $\delta = \delta'$. Per això, parlarem de mitjanes d'ingressos encara que també podríem considerar els ingressos totals.

Altres indicadors de desigualtat, com ara els emprats per l'OCDE (OECD 2023) fan servir no només la comparació dels ingressos o les mitjanes d'un quintil, sinó el punt de tall entre un decil i altre, allò que podríem dir els *interdecils*. Així s'han formulat diversos indicadors: la relació entre els interdecils del decil superior i l'inferior, el que es podria formular com: $P90/P10$ (és clar que el punt de tall del decil superior és el valor del percentil 90 i així successivament); la relació entre l'*interdecil* superior i el mitjà, val a dir, el valor de la mitjana, que seria: $P90/P50$, i el de la mitjana amb l'*interdecil* inferior, és a dir: $P50/P10$. L'OCDE també calcula $S90/S40$, altrament dit la proporció entre els ingressos del decil superior i els dels dos quintils inferiors (o les seues mitjanes, tant s'hi val, com hem explicat adés), el que s'anomena Índex de Palma).

La comparació de quantils, bé es considere S o bé es considere P , o fins i tot les altres possibilitats mostrades, es susceptible de rebre dos crítiques, que ens semblen molt importants.

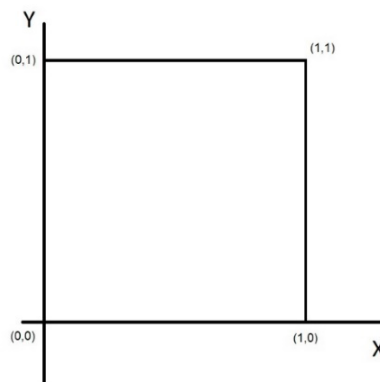
En primer lloc, els procediment bandeja altres quantils (per exemple, $S80/S20$ desconsidera el 60% de la població, és a dir, la que forma part dels quintils 2n, 3r i 4t). Per això, un mateix valor de δ pot correspondre a situacions ben diferents, segons els valors dels quintils bandejats. En segon lloc, la comparació de quantils ignora el fenomen dels mega-rics (Piketty 2013; també Piketty 2023). Aquestes dues crítiques no són aplicables a l'Índex de Gini.

L'Índex de Gini és la mesura de desigualtat social estandarditzada (vegeu, per exemple,

Banc Mundial 2024, INE 2023).

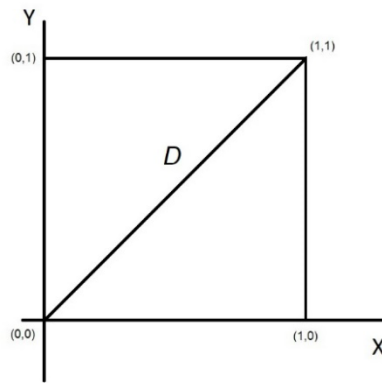
L'Índex de Gini (que pren el seu nom de l'economista italià Corrado Gini) parteix del fet que una representació en els eixos cartesianes del percentatge *acumulat* de la població (en l'eix de les abscisses, X) i del percentatge *acumulat* d'un recurs econòmic –com ara els ingressos– (en l'eix de les ordenades, Y), que són quantitats que oscil·len entre el 0,0% i el 100,0% (o en expressió decimal entre 0,0 i 1,0), esdevé sempre en el que anomenarem un *quadrat unitat* perquè és el quadrat que té una unitat de costat, que està definit pels punts: (0, 0), (1, 0), (1, 1) i (0, 1) (vegeu Gràfic 1).

GRÀFIC 1. QUADRAT UNITAT

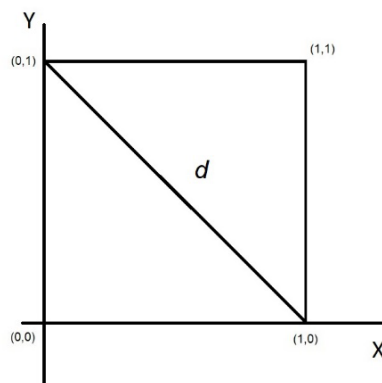


Com que el costat d'aquest quadrat mesura 1, per definició, la seua àrea (base $\times$ altura) és també 1 (una unitat gràfica). Al quadrat unitat es poden traçar dos diagonals: la que uneix els punts (0, 0) i (1, 1), que anomenarem *diagonal creixent* (i abreviarem D) (vegeu el gràfic 2), i la que uneix els punts (0, 1) i (1, 0), que anomenarem *diagonal decreixent* (i abreviarem d) (vegeu el gràfic 3).

GRÀFIC 2. DIAGONAL CREIXENT



GRÀFIC 3. DIAGONAL DECREIXENT

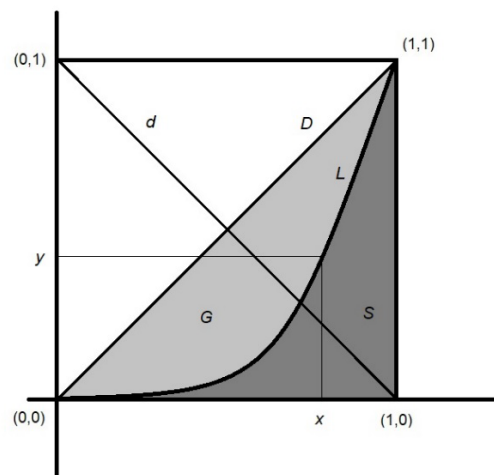


Els adjectius *creixent* i *decreixent* es refereixen, en definitiva, a la convenció que els eixos cartesianes presenten quantitats superiors a la dreta i inferiors a l'esquerra (i dalt i baix, respectivament), la qual cosa està ancorada en la manera com llegim d'esquerra a dreta, el que no és una pràctica universal. És fàcil deduir que D i d són hipotenuses de triangles rectangles isòsceles, els catets dels quals mesuren 1 (per definició), per la qual cosa la llargària d'aquestes hipotenuses (les diagonals) serà igual a $\sqrt{2}$ (per l'aplicació del teorema de Pitàgores). També és trivial que aquests triangles tenen una superfície de $\frac{1}{2}$ unitat gràfica.

La premissa bàsica de l'Índex de Gini és que, en una societat plenament igualitària, la representació en els eixos cartesianes del percentatge acumulat de la població (en l'eix X) i del percentatge acumulat d'un recurs econòmic –com ara els ingressos– (en l'eix Y) se superposaria a la diagonal creixent, D . És fàcil d'entendre-ho: en una societat igualitària el 10% dels individus tindrien el 10% dels recursos; el 20% dels individus, el 20% dels recursos, etc. En una societat no igualitària, la representació dels punts no és la diagonal

D , sinó una corba (L , l'anomenada corba de Lorenz) per baix d'ella¹, però que coincideix amb la diagonal necessàriament en els punts $(0, 0)$, (perquè el 0% de població acumulada disposa necessàriament del 0% dels recursos) i pel punt $(1, 1)$ (perquè el 100% dels recursos acumulats correspon al 100% de la població). Com menys igualitària (o com més desigualitària) siga la societat, més s'allunyarà la corba de la diagonal creixent, i a l'inrevés. La proporció entre la superfície tancada per la diagonal D i la corba L , superfície que en el gràfic posterior anomenem G , i tota la superfície per baix de la diagonal, que s'ha marcat com a S , proporciona un coeficient, que multiplicat per 100 dona l'Índex de Gini. Partim del gràfic 4.

GRÀFIC 4. REPRESENTACIÓ DEL CÀLCUL DE L'ÍNDEX DE GINI



S'hi pot veure com és fàcil calcular S (el clàssic problema de superfície baix de corba, que es resol amb una integral definida de la funció de la corba, com s'explicarà més endavant) i, a partir del seu valor, establir el de G .

$$G = \frac{1}{2} - S$$

Es calcula el Coeficient de Gini (CG) de la manera següent:

¹ Deixem de banda algun cas, com ara, la representació de la distribució dels recursos socials redistributius, com ara les pensions de jubilació o els subsidis d'ocupació, en els quals aniria per dalt de la diagonal. En aquests casos, aquests recursos representen una tensió cap a dalt la corba del mercat.

$$CG = \frac{G}{G+S} = \frac{G}{\frac{1}{2}} = 2G$$

I després es defineix l'Índex de Gini (IG), de manera que:

$$IG = 100 CG$$

La genialitat de Gini és que, si estem en el quadrat unitat, aquest càlcul se simplifica considerablement, com mostrarem a continuació.

Suposem una funció de la corba realment complicada, com puga ser la representada per una equació polinòmica de sisé grau (val a dir, que presenta com a màxim exponent de la incògnita un valor de x^6 , etc.) la seua integral és d'un grau superior (x^7 , etc.), però com que el quadrat unitat té com a valors extrems de les abscisses els punts 0 i 1, el càlcul se simplifica extraordinàriament i es redueix a una suma de fraccions. Ho formularem en terminologia matemàtica.

Suposem que l'equació de la corba té la forma de l'equació polinòmica de grau sis:

$$y = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$$

Aleshores, l'àrea baix de la corba es calcula mitjançant la integral definida de la funció entre els punts 0 i 1. Val a dir:

$$\begin{aligned} S &= \int_0^1 (ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g) dx = \\ &= a \int_0^1 x^6 dx + b \int_0^1 x^5 dx + c \int_0^1 x^4 dx + d \int_0^1 x^3 dx + e \int_0^1 x^2 dx + f \int_0^1 x dx \\ &\quad + g \int_0^1 dx = \\ &= a \left[\frac{x^7}{7} \right]_0^1 + b \left[\frac{x^6}{6} \right]_0^1 + c \left[\frac{x^5}{5} \right]_0^1 + d \left[\frac{x^4}{4} \right]_0^1 + e \left[\frac{x^3}{3} \right]_0^1 + f \left[\frac{x^2}{2} \right]_0^1 + g [x]_0^1 = \\ &= \frac{a}{7} + \frac{b}{6} + \frac{c}{5} + \frac{d}{4} + \frac{e}{3} + \frac{f}{2} + g \end{aligned}$$

Com s'hi pot veure, si sabem la funció polinòmica (de l'ordre que siga), el càlcul *es resol amb operacions relativament senzilles* (una suma de fraccions, en aquest cas). Aquesta simplificació és possible, com hem dit, precisament perquè la integral està definida entre

els punts 0 (que anul·la el polinomi en realitzar la substitució de x per 0: qualsevol número elevat a 0 és 0) i 1 (1 elevat a qualsevol potència és 1). Aquesta és la raó de fer servir percentatges acumulats.

Potser per aquesta senzillesa, l'Índex de Gini s'ha estès en molts àmbits científics, més enllà dels estudi de les desigualtats socioeconòmiques. Per citar alguns exemples ben heterogenis: l'aqüicultura (Pérez et al. 2005), la psicopatologia (Maldonado et al. 2007), la pluviometria (Benhamrouche et al. 2012) o la distribució botànica (Grajales et al. 2016), però no ens consta que siga el cas en la sociolingüística o la sociologia de la llengua.

Pot donar-se el cas que disposem només d'un punt en el quadrat unitat, i més concretament, com s'ha explicat, en la meitat per baix de la diagonal creixent D , que anomenarem el punt (x_i, y_i) . En realitat, disposem de dos punts més en cada cas, perquè, com he dit adés, cal tindre en compte els punts $(0, 0)$ i $(1, 1)$. Naturalment, hi ha infinites corbes que passen pels punts $(0, 0)$, (x_i, y_i) , $(1, 1)$. *Cada tipus de corba correspondrà a un model diferent.* Tanmateix, podem fer servir aquelles *més simples* en la seua caracterització. fent servir les corbes més senzilles, en el nostre cas les corbes exponencials i les corbes parabòliques.

Les corbes exponencials corresponen a funcions exponencials. A tall d'exemple senzill:

$$y = 2^x$$

On la variable independent és un exponent.

Les corbes parabòliques estan definides per funcions com ara:

$$y = x^2$$

I lògicament són simètriques, perquè:

$$x^2 = (x)^2 = (-x)^2$$

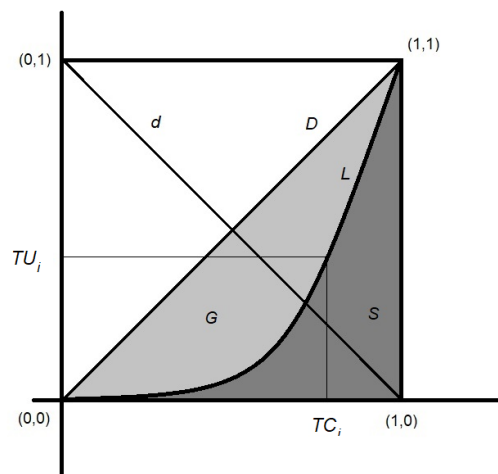
Si x i $-x$, al quadrat, proporcionen el mateix valor de y , aleshores la corba resultant serà simètrica, en aquest exemple, respecte de l'eix de les ordenades (encara que podem variar l'eix de simetria, com veurem).

2.- Adaptació de l'Índex de Gini en un Índex de Desigualtat Lingüística

D'aquesta manera, com ha estat defensat (Hernández 2020), és possible aplicar la metodologia de l'Índex de Gini per tal d'elaborar un Índex de Desigualtat Lingüística. Partim d'una taxa de competència (TC) i una taxa d'ús (TU) i entenem que la desigualtat és la coacció (en sentit de Durkheim) que s'estableix quan la competència no pot ser realitzada pel mitjà de l'ús. Aquesta no realització és la que ens permet comparar diverses societats. Així, per exemple, si suposem dues societats, N_1 i N_2 , i per a cadascuna d'elles disposem d'un punt (una parella de valors) de la mateixa sèrie, val a dir, de dos parells de valors (TC_1, TU_1) i (TC_2, TU_2) respectivament, podem calcular els Índex de Desigualtat Lingüística corresponents i comparar-les.

Comencem amb la *corba exponencial*, tal com es representa en l'anterior gràfic 4, que tornarem a copiar, referit ara a l'Índex de Desigualtat Lingüística (gràfic 5).

GRÀFIC 5. REPRESENTACIÓ DEL CàLCUL DE L'ÍNDEX DE DESIGUALTAT LINGÜÍSTICA AMB UNA CORBA EXPONENCIAL



De les corbes exponencials que aconsegueixen la condició de passar pels punts $(0, 0)$, (TC_i, TU_i) i $(1, 1)$, considerarem una de ben senzilla:

$$y = (2^x - 1)^z$$

Perquè:

$$\text{si } x = 0, \text{ aleshores } (2^0 - 1)^z = (1 - 1)^z = 0^z = 0 = y$$

$$\text{si } y = 1, \text{ aleshores } (2^x - 1)^z = 1; 2^x - 1 = \sqrt[z]{1}; 2^x - 1 = 1; 2^x = 2; 2^x = 2^1; x = 1$$

Per a esbrinar el valor de z procedirem a calcular el logaritme (decimal o en base 10) dels dos membres de l'equació:

$$\text{si } y = (2^x - 1)^z, \text{ aleshores } \log y = \log (2^x - 1)^z = z \log (2^x - 1)$$

I dividint els dos membres de la igualtat per « $\log(2^x - 1)$ » o, més vulgarment, aïllant z :

$$z = \frac{\log y}{\log(2^x - 1)}$$

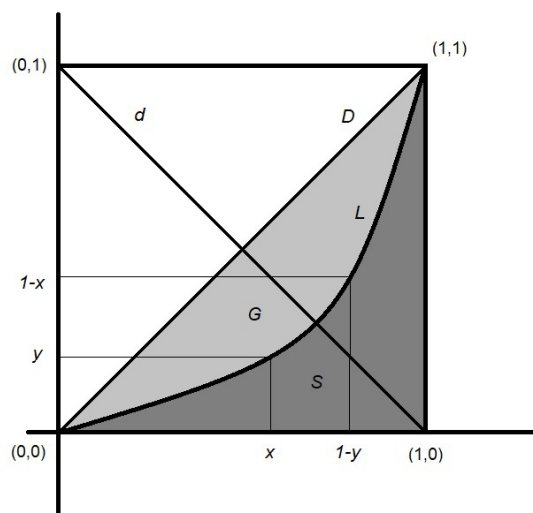
Val a dir, per al nostre cas:

$$z = \frac{\log TU}{\log(2^{TC} - 1)}$$

Una vegada es disposa del valor z , es pot fer la corba d'una manera senzilla. S'assignen valors entre 0,00 i 1,00 a x (TC) en l'expressió $y = (2^x - 1)^z$, y s'obtenen valors de y (TU). Aquests valors es representen en uns eixos cartesianes per mitjà d'un full de càlcul que ens pot proporcionar la corresponent equació polinòmic del màxim grau que permeti el programa. A continuació es calcula la integral definida per esbrinar la superfície S i, posteriorment, la superfície G , pel procediment ja comentat. Aquest és el procediment que es mostrarà més endavant.

També es pot calcular l'Índex de Desigualtat Lingüística amb el model de la corba parabòlica, que presenta la condició de simetria respecte de la diagonal menor d , com mostra el gràfic 6.

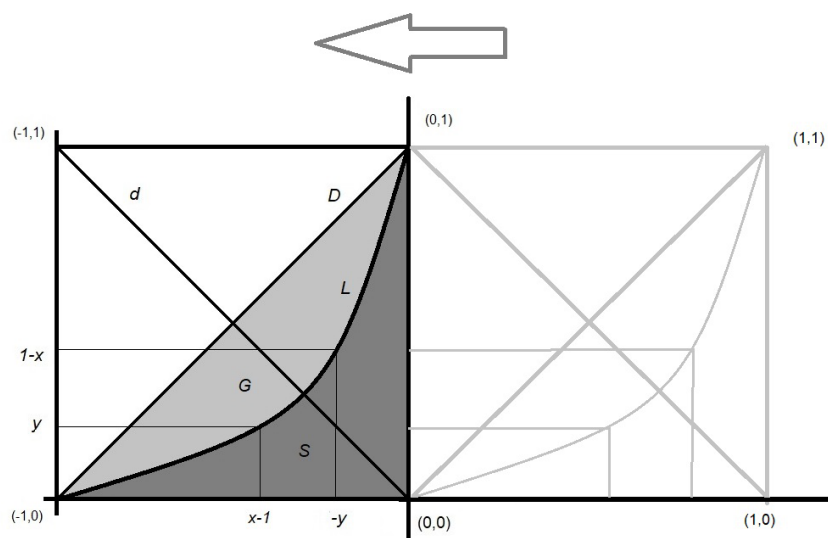
GRÀFIC 6. MODEL DE LA CORBA PARABÒLICA



La condició de simetria pot semblar exagerada; però no ho és, si considerem l'extensió de la simetria en la natura (vegeu Navarro 2011). Per tal d'esbrinar la funció de la corba hem de procedir a certes transformacions. Doncs bé, si desplaçem els eixos cartesianos fins a fer coincidir la diagonal decreixent amb l'eix d'ordenades, podrem fàcilment demanar-li a un programa (full de càlcul) que calcule la funció de la corba parabòlica que passa pels punts $(0, 0)$, (x, y) , $(1 - y, 1 - x)$ i $(1, 1)$. Explicarem amb detall aquest procediment.

Partim del gràfic 6. En primer lloc realitzarem una translació, val a dir, un desplaçament horitzontal de l'eix Y (o del quadrat unitari, és igual) la qual cosa es pot representar en el gràfic 7.

GRÀFIC 7. TRANSLACIÓ D'EIXOS CARTESIANS



És el mateix desplaçar la figura cap a l'esquerra que desplaçar els eixos cap a la dreta.

Aquesta translació produeix unes noves coordenades en el quadrat unitat. D'aquesta manera els punts $(0, 0)$, (x, y) , $(1 - y, 1 - x)$ i $(1, 1)$ del gràfic 6 es transformen en uns altres, que s'arreglen en la taula 1: Els valors de x disminueixen en 1, mentre que els de y es mantenen iguals. Altrament dit, si x' (ací x' vol dir: x prima; no res a veure amb derivades) i y' són les noves coordenades, s'acompleix que:

TAULA 1. COORDENADES ORIGINALS I COORDENADES DESPRÉS DE LA TRANSLACIÓ

Coordenades originals (x, y)	Coordenades després de la translació (x', y')
$(0, 0)$	$(-1, 0)$
$(0, 1)$	$(-1, 1)$
$(1, 0)$	$(0, 0)$
$(1, 1)$	$(0, 1)$
(x, y)	$(x - 1, y)$
$(1 - y, 1 - x)$	$(-y, 1 - x)$

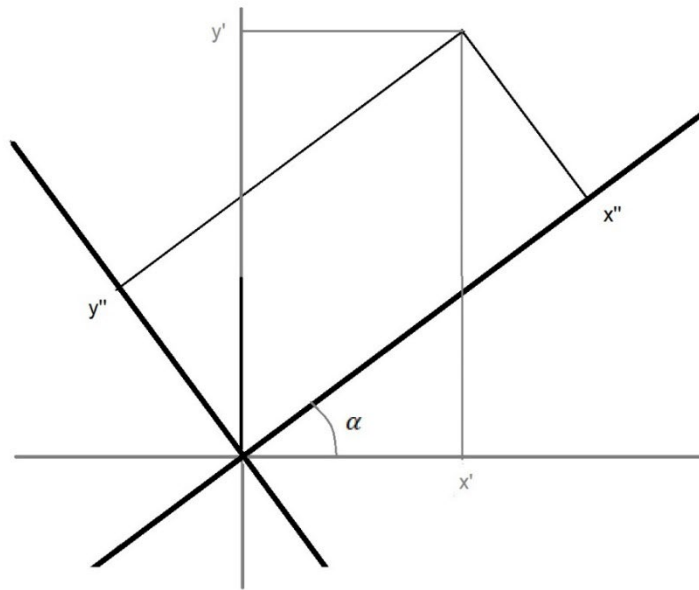
Font: Elaboració pròpia.

A continuació efectuarem una rotació En primer lloc, explicarem el principi general; després, l'aplicarem al nostre cas.

Suposem que tenim uns eixos cartesianes amb un punt que té unes coordenades (x', y') (és clar que ja els hem anomenat « x', y' », i no simplement « x, y », per donar a entendre que ja estem treballant amb les coordenades després de la translació anterior). Si fem una

rotació dels eixos amb un angle α en el sentit contrari de les agulles del rellotge, el punt (x', y') adoptarà unes noves coordenades, que anomenarem (x'', y'') , tal com mostra el gràfic 8.

GRÀFIC 8. ROTACIÓ D'EIXOS CARTESIANS. REPRESENTACIÓ GENERAL



En el cas de les rotacions, les noves coordenades (x'', y'') estan establides per les fórmules següents (vegeu: Aleksandrov, Kolmogorov & Lavrentiev 2015; Pontriaguin 2011a i 2011b.), on α és l'angle de rotació.

$$x'' = x' \cos \alpha + y' \sin \alpha$$

$$y'' = y' \cos \alpha - x' \sin \alpha$$

Doncs bé, en el nostre cas anem a imprimir una rotació de 45° (en el sentit de les agulles d'un rellotge analògic –o de la seua simulació digital). Podem expressar la transformació en funció de les coordenades originals. Així obtindrem:

$$x'' = (x - 1) \cos \alpha + y \sin \alpha$$

$$y'' = y \cos \alpha - (x - 1) \sin \alpha$$

En aquest cas, com que es tracta d'una rotació de 45° , tenim que²:

$$\sin 45^\circ = \cos 45^\circ = \frac{\sqrt{2}}{2}$$

Per la qual cosa, substituint aquest valor en les fórmules anteriors, obtindrem:

$$x'' = \left(\frac{\sqrt{2}}{2}\right)(x - 1 + y) = \frac{(x - 1 + y)\sqrt{2}}{2}$$

$$y'' = \left(\frac{\sqrt{2}}{2}\right)(y + 1 - x) = \frac{(y + 1 - x)\sqrt{2}}{2}$$

Aleshores, les noves coordenades (x'', y'') prenen en el nostre cas els valors que apareixen en la taula 2 i el gràfic 7 pren la forma del gràfic 9:

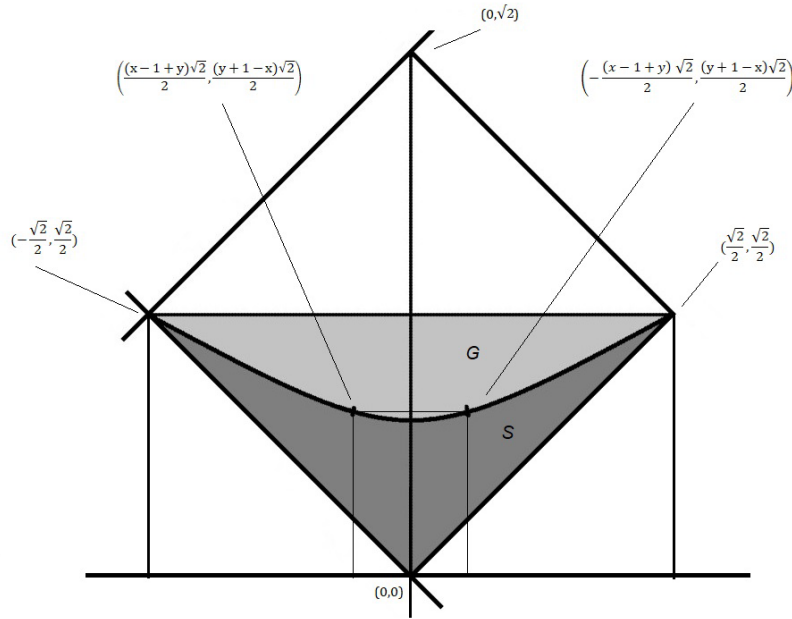
TAULA 2. COORDENADES DESPRÉS DE LA TRANSLACIÓ I LA ROTACIÓ

Coordenades originals (x, y)	Coordenades després de la translació (x', y')	Coordenades després de la translació i la rotació (x'', y'')
(0, 0)	(-1, 0)	$(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$
(0, 1)	(-1, 1)	$(0, \sqrt{2})$
(1, 0)	(0, 0)	(0, 0)
(1, 1)	(0, 1)	$(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$
(x, y)	(x - 1, y)	$(\frac{(x - 1 + y)\sqrt{2}}{2}, \frac{(y + 1 - x)\sqrt{2}}{2})$

² Cal recordar que el *sinus* i el *cosinus* són raons trigonomètriques, val a dir, proporcions entre els catets i la hipotenusa d'un triangle rectangle. En el cas del *sinus* es defineix com la proporció entre el catet oposat a l'angle recte i la hipotenusa; en el cas del *cosinus* com la proporció entre el catet contigu a l'angle recte i la hipotenusa. En el cas d'un angle de 45° , la proporció entre catet oposat a l'angle recte i hipotenusa és de $\frac{\sqrt{2}}{2}$. Com que la suma dels angles d'un triangle pla és 180° , cas d'un triangle rectangle amb un angle de 45° , l'altre angle ha de mesurar necessàriament 45° , per la qual cosa el catet oposat del primer angle és el catet oposat del segon i viceversa; és per això que en aquest cas, el sinus és igual al cosinus. En el cercle trigonomètric (de radi 1 i, per tant, d'hipotenusa 1), el sinus i el cosinus d'un angle de 45° mesuren exactament $\frac{\sqrt{2}}{2}$ unitats gràfiques, el que és equivalent a 0,707106...

$(1 - y, 1 - x)$	$(-y, 1 - x)$	$\left(-\frac{(x - 1 + y)\sqrt{2}}{2}, \frac{(y + 1 - x)\sqrt{2}}{2}\right)$
------------------	---------------	--

GRÀFIC 9. ROTACIÓ D'EIXOS CARTESIANS, REPRESENTACIÓ PER AL NOSTRE CAS



Aleshores es tracta de calcular la funció d'una corba que passa pels punts $\left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$, $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$ i $\left(\frac{(x-1+y)\sqrt{2}}{2}, \frac{(y+1-x)\sqrt{2}}{2}\right)$. En la taula 2 anterior també hem afegit el simètric de (x, y) respecte de la diagonal decreixent, és a dir $(-y, 1 - x)$, que, realitzada la translació i la rotació, pren el valor $\left(-\frac{(x-1+y)\sqrt{2}}{2}, \frac{(y+1-x)\sqrt{2}}{2}\right)$. Després de disposar de la funció de la corba i per tal de calcular la superfície baix de la corba, cal fer la integral definida entre $-\frac{\sqrt{2}}{2}$ i $\frac{\sqrt{2}}{2}$. A continuació cal restar la superfície dels dos triangles de base i altura $\frac{\sqrt{2}}{2}$, que tenen en conjunt àrea $\frac{1}{2}$, ja que

$$2 \left(\frac{\frac{\sqrt{2}}{2} \frac{\sqrt{2}}{2}}{2} \right) = \frac{1}{2}$$

Pot semblar complicat, per la proliferació de fórmules amb arrels quadrades, però realment no ho és.

Com que la superfície del quadrat té com a base és el segment que va des del valor $-\frac{\sqrt{2}}{2}$ al valor $\frac{\sqrt{2}}{2}$ de l'eix X i com a alçària el segment que va des del valor 0,0 al valor $\frac{\sqrt{2}}{2}$ de l'eix Y , el quadrat té una superfície d'una unitat³, aleshores:

$$\begin{aligned} G &= 1 - \left[\int_{-\frac{\sqrt{2}}{2}}^{\frac{\sqrt{2}}{2}} (ax^2 + bx + c) dx \right] = 1 - 2 \left[\int_0^{\frac{\sqrt{2}}{2}} (ax^2 + bx + c) dx \right] \\ &= 1 - 2 \left[a \left[\frac{x^3}{3} \right]_0^{\frac{\sqrt{2}}{2}} + b \left[\frac{x^2}{2} \right]_0^{\frac{\sqrt{2}}{2}} + c [x]_0^{\frac{\sqrt{2}}{2}} \right] = 1 - 2 \left[a \left[\frac{x^3}{3} \right]_0^{\frac{\sqrt{2}}{2}} + c [x]_0^{\frac{\sqrt{2}}{2}} \right] \\ &= 1 - 2 \left[\frac{\left(\frac{\sqrt{2}}{2} \right)^3}{3} a + \frac{\sqrt{2}}{2} c \right] = 1 - \left[\frac{\sqrt{2}}{6} a + \sqrt{2} c \right] \end{aligned}$$

I per tant:

$$CG = 2G = 2 \left\{ 1 - \left[\frac{\sqrt{2}}{6} a + \sqrt{2} c \right] \right\} = 2 - \frac{\sqrt{2}}{3} a - 2\sqrt{2}c$$

Per la qual cosa podem formular l'Índex de Desigualtat Lingüística segons un model de corba parabòlica així:

$$\begin{aligned} IDL &= 100CG = 100(2G) = 200 \left[1 - 2 \int_0^{\frac{\sqrt{2}}{2}} (ax^2 + bx + c) dx \right] = 200 \left\{ 1 - \right. \\ &\left. 2 \left(a \left[\frac{x^3}{3} \right]_0^{\frac{\sqrt{2}}{2}} + b \left[\frac{x^2}{2} \right]_0^{\frac{\sqrt{2}}{2}} + c [x]_0^{\frac{\sqrt{2}}{2}} \right) \right\} = 200 \left[1 - \left(\frac{a\sqrt{2}}{6} + b + c\sqrt{2} \right) \right] \end{aligned}$$

Com que b és pràcticament zero (en tractar-se d'una corba parabòlica), es pot simplificar en:

³ És senzill, perquè l'àrea (A) és base ($2 \frac{\sqrt{2}}{2}$) per alçària ($\frac{\sqrt{2}}{2}$), aleshores: $A = 2 \frac{\sqrt{2}}{2} \times \frac{\sqrt{2}}{2} = \sqrt{2} \times \frac{\sqrt{2}}{2} = \frac{2}{2} = 1$.

$$IDL = 200 \left(1 - \frac{a\sqrt{2}}{6} - c\sqrt{2} \right) = 200 - \frac{200a\sqrt{2}}{6} - 200c\sqrt{2}$$

Que representa la fórmula del càlcul de l'índex de desigualtat lingüística (segons el model d'una corba parabòlica).

3.- Les dades de les regions sociolingüístiques valencianes

En les enquestes sociolingüístiques, la zona valencianoparlant es divideix en les regions que s'indiquen en el requadre 1.

REQUADRE 1. REGIONS SOCIOLINGÜÍSTIQUES DE LA ZONA VALENCIANOPARLANT

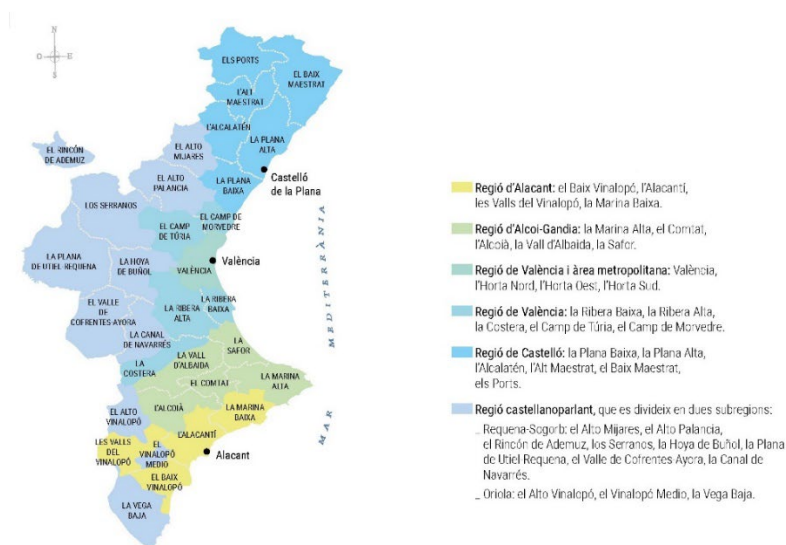
Regió d'Alacant (valencianoparlant): comarques del Baix Vinalopó, l'Alacantí, les Valls del Vinalopó i la Marina Baixa.

Regió d'Alcoi-Gandia: comarques de la Marina Alta, el Comtat, l'Alcoià, la Vall d'Albaida i la Safor.

Regió de València i l'àrea metropolitana: ciutat de València, l'Horta Nord, l'Horta Oest i l'Horta Sud.

Regió de València (valencianoparlant): comarques de la Ribera Baixa, la Ribera Alta, la Costera, el Camp de Túria i el Camp de Morvedre.

Regió de Castelló (valencianoparlant): comarques de la Plana Baixa, la Plana Alta, l'Alcalatén, l'Alt Maestrat, el Baix Maestrat i els Ports.



Es fan servir les Enquestes de Coneixement i Ús del Valencià realitzades per la Conselleria d'Educació (amb distintes denominacions en el període) entre 2005 i 2021 (EUCV 2005, 2010, 2015 i 2021). En el cas de les competències (oral activa: val a dir, el percentatge de població que afirma que sap parlar valencià), les respostes possibles són: *perfectament, bastant bé, un poc i gens*, a les freqüències de les respostes de les quals s'apliquen uns coeficients de ponderació de 3/3, 2/3, 1/3 i 0/3. En el cas dels usos lingüístics (ús al carrer amb persones desconegudes), les respostes possibles són: *sempre en valencià, generalment en valencià, més valencià que castellà, indistintament, més castellà que valencià, generalment en castellà, sempre en castellà i altres llengües*. Els coeficients de ponderació aplicats són: 6/6, 5/6, 4/6, 3/6, 2/6, 1/6 i 0/6 per als dos últims casos. En ambdós casos es depura la base de dades, amb una distribució ponderada dels valors ns/nc, que són ínfims (amb la finalitat de fer una aplicació més acurada de l'algorisme que s'explicarà més endavant).

Model de corba exponencial

A continuació s'explica amb detall el procediment en un cas i es proporcionen les dades de la resta.

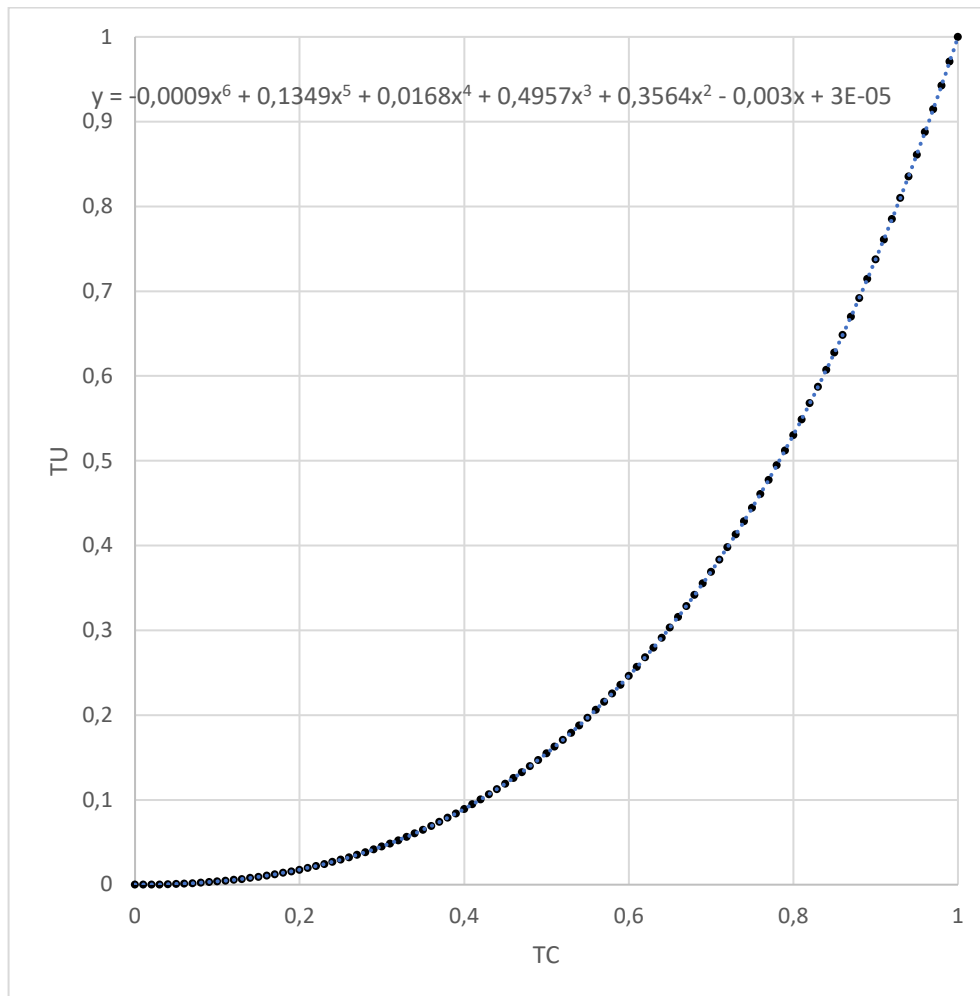
Considerem les dades de la regió sociolingüística de la regió d'Alacant valencianoparlant. Els resultats de les freqüències de les respostes d'EUCV 2005 per a les *TC* (oral activa) i *TU* (al carrer, amb persones desconegudes) són, respectivament: 0,454 i 0,122 (o 45,4% i 12,2% en expressió percentual). Cal advertir que ací i en les taules les dades estan arrodonides o es presenten fins al cinqué decimal, encara que s'han fet servir en els càlculs tots els decimals possibles. Se substitueixen aquestes dades en la fórmula:

$$z = \frac{\log TU}{\log(2^{TC} - 1)}$$

i es calcula z , que presenta un valor de: 2,11631. Aleshores es donen valors des de 0,00 a 1,00 en la fórmula de la corba exponencial $y = (2^x - 1)^z$, que en aquest cas és: $TU =$

$(2^{TC} - 1)^{2,1163}$, i es compona la corba del gràfic 10:

GRÀFIC 10. REPRESENTACIÓ DE LA CORBA EXPONENCIAL (ALACANT, EUCV 2005)



Al gràfic s'indica la línia de tendència polinòmica que passa pels punts i la seua equació⁴. Com hem mostrat més amunt, si tenim una equació de la corba del tipus:

$$ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$$

La superfície baix de la corba S entre els punts $x = 0$ i $x = 1$ es pot calcular fàcilment com:

⁴ S'hi fa servir la notació abreujada, per la qual $3E - 05 = 3 \times 10^{-5} = 3/10^5 = 0,00003$.

$$S = \frac{a}{7} + \frac{b}{6} + \frac{c}{5} + \frac{d}{4} + \frac{e}{3} + \frac{f}{2} + g$$

En aquest cas:

$$S = -0,00013 + 0,02248 + 0,00336 + 0,12393 + 0,11880 - 0,00150 + 3,00E-05 = 0,26696$$

I per les fórmules anteriors:

$$G = 0,23303$$

$$CG = 0,46606$$

$$IDL: 46,60604$$

En la taula 3, a continuació, s'arreglen les dades de les regions i enquestes esmentades, que s'han calculat repetint el procediment indicat.

TAULA 3. IDL PER A LES REGIONS I ENQUESTES (CORBA EXPONENCIAL)

	<i>TC ponderada</i>	<i>TU ponderada</i>	<i>z</i>	<i>IDL</i>
Regió d'Alacant (valencianoparlant)				
EUCV 2005	0,45400	0,12183	2,11631	46,60604
EUCV 2010	0,39700	0,12583	1,80308	39,82538
EUCV 2015	0,50833	0,13916	2,28883	49,71985
EUCV 2021	0,37266	0,07716	2,09699	46,23176
Regió d'Alcoi-Gandia				
EUCV 2005	0,81366	0,71000	1,23418	22,70528
EUCV 2010	0,71900	0,60350	1,15591	18,80504
EUCV 2015	0,75600	0,60733	1,33764	26,01961
EUCV 2021	0,63800	0,49783	1,18887	20,22042
Regió de València i l'àrea metropolitana				
EUCV 2005	0,70733	0,27616	2,81178	57,33723
EUCV 2010	0,64100	0,22717	2,55145	53,85933
EUCV 2015	0,67666	0,28416	2,45058	52,35147
EUCV 2021	0,60900	0,17300	2,72443	56,23366
Regió de València (valencianoparlant)				

EUCV 2005	0,79066	0,70216	1,12290	17,35980
EUCV 2010	0,75066	0,63783	1,17750	19,74580
EUCV 2015	0,75666	0,62716	1,25526	22,91561
EUCV 2021	0,68333	0,46166	1,54233	32,78304
Regió de Castelló (valencianoparlant)				
EUCV 2005	0,70733	0,41500	1,92181	42,58800
EUCV 2010	0,64100	0,35600	1,77805	39,21057
EUCV 2015	0,67666	0,37100	1,93125	43,43723
EUCV 2021	0,60900	0,26100	2,08585	46,01442

Font: Elaboració pròpia a partir de EUCV 2005, 2010, 2015 i 2021.

Model de corba parabòlica

Explicarem el càlculs per al primer cas i proporcionarem les dades de la resta.

En el cas de la regió d'Alacant valencianoparlant, en l'any 2005, els valors de TC i TU foren, respectivament: 0,45400 i 0,12183. Amb aquestes dades, la taula transformada de coordenades és:

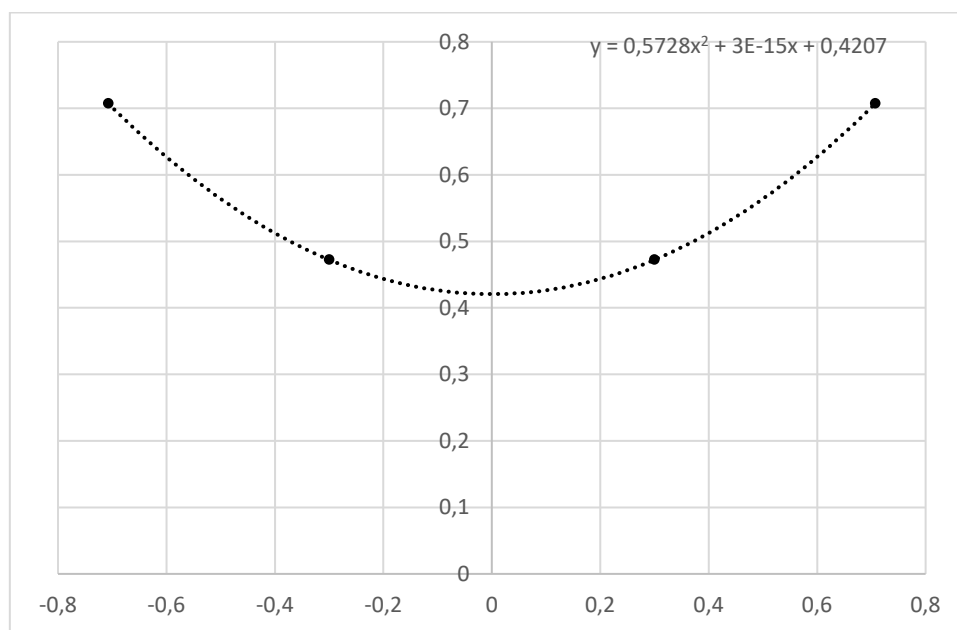
TAULA 4. COORDENADES DESPRÉS DE LA TRANSLACIÓ I LA ROTACIÓ EN EL CAS

Coordenades originals (x, y)	Coordenades després de la translació i la rotació (x'', y'')	Valors en aquest cas
(0, 0)	$(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$	(-0,70711, 0,70711)
(1, 1)	$(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$	(0,70711, 0,70711)
(x, y)	$(\frac{(x-1+y)\sqrt{2}}{2}, \frac{(y+1-x)\sqrt{2}}{2})$	(-0,29993, 0,47222)
(1-y, 1-x)	$(-\frac{(x-1+y)\sqrt{2}}{2}, \frac{(y+1-x)\sqrt{2}}{2})$	(0,29993, 0,47222)

Nota: Com sempre, els càlculs es realitzen amb més decimals dels que consten a les taules.

La qual cosa proporciona la representació gràfica següent (amb la línia de tendència polinòmica de grau 2 i la seua equació):

GRÀFIC 11. REPRESENTACIÓ DE LA CORBA PARABÒLICA (ALACANT, EUCV 2005)



Font: Elaboració pròpia amb les dades de la taula 4.

S'hi aplica la fórmula anterior i s'obté el valor de l'*IDL*.

$$IDL = 200 \left(1 - \frac{a\sqrt{2}}{6} - c\sqrt{2} \right) = 200 - \frac{200a\sqrt{2}}{6} - 200c\sqrt{2} = 200 - \frac{200(0,5728)\sqrt{2}}{6} - 200(0,4207)\sqrt{2} = 54,00602$$

En la taula 5, a continuació, s'arreglen les dades de les regions i enquestes esmentades, que s'han calculat repetint el procediment indicat (també s'hi fa constar els valors dels paràmetres *a* i *c* de l'equació corresponent de segon grau).

TAULA 5. IDL PER A LES REGIONS I ENQUESTES (CORBA PARABÒLICA)

	<i>TC ponderada</i>	<i>TU ponderada</i>	<i>Paràmetre a</i>	<i>Paràmetre c</i>	<i>IDL</i>
Regió d'Alacant (valencianoparlant)					

EUCV 2005	0,45400	0,12183	0,5728	0,4207	54,00602
EUCV 2010	0,39700	0,12583	0,4965	0,4588	46,82653
EUCV 2015	0,50833	0,13916	0,5962	0,4090	56,21219
EUCV 2021	0,37266	0,07716	0,5993	0,4075	56,49032
Regió d'Alcoi-Gandia					
EUCV 2005	0,81366	0,71000	0,202	0,6061	19,04666
EUCV 2010	0,71900	0,60350	0,1823	0,6160	17,17518
EUCV 2015	0,75600	0,60733	0,2422	0,5860	22,83675
EUCV 2021	0,63800	0,49783	0,2020	0,6061	19,04666
Regió de València i l'àrea metropolitana					
EUCV 2005	0,70733	0,27616	0,6099	0,4021	57,51798
EUCV 2010	0,64100	0,22717	0,5956	0,4093	56,15562
EUCV 2015	0,67666	0,28416	0,5559	0,4291	52,42681
EUCV 2021	0,60900	0,17300	0,6474	0,3834	61,03938
Regió de València (valencianoparlant)					
EUCV 2005	0,79066	0,70216	0,1653	0,6245	15,57241
EUCV 2010	0,75066	0,63783	0,1879	0,6131	17,73144
EUCV 2015	0,75666	0,62716	0,2148	0,5997	20,25346
EUCV 2021	0,68333	0,46166	0,3202	0,5470	30,19066
Regió de Castelló (valencianoparlant)					
EUCV 2005	0,70733	0,41500	0,4197	0,4973	39,55747
EUCV 2010	0,64100	0,35600	0,4031	0,5056	37,99241
EUCV 2015	0,67666	0,37100	0,4333	0,4905	40,83969
EUCV 2021	0,60900	0,26100	0,5006	0,4568	47,19894

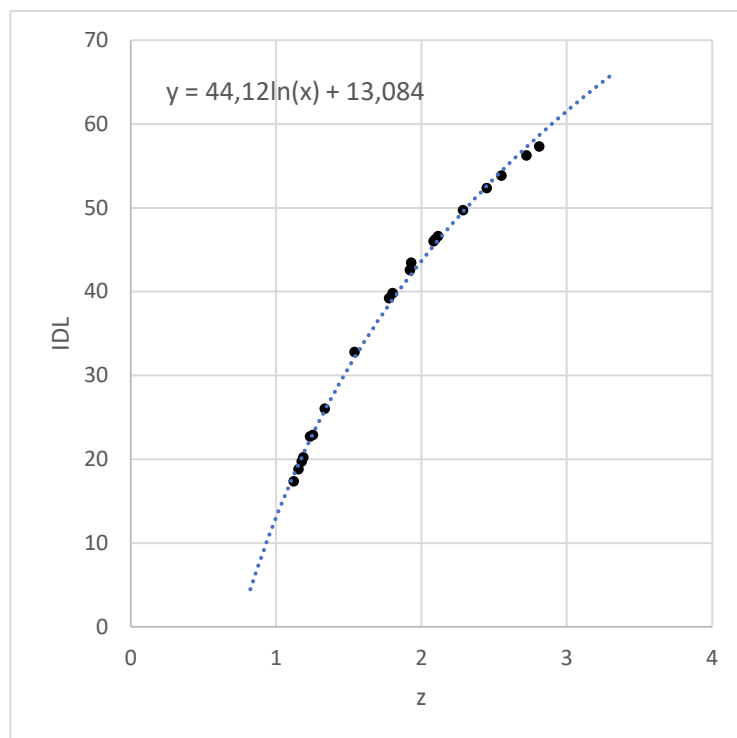
Font: Elaboració pròpia a partir de EUCV 2005, 2010, 2015 i 2021.

4.- Ampliació i debat

El càlcul anterior de l'Índex de Desigualtat Lingüística (*IDL*) permet dues ampliacions teòriques per a futures recerques. En primer lloc, la possibilitat de fer una fórmula que relaciones les dades de *TC* i *TU* directament amb una aproximació a l'*IDL*; en segon lloc, el debat al voltant de l'anomenada Corba del Gran Gatsby.

En primer lloc, podem realitzar una aproximació als valors de l'*IDL* a partir dels valors de *z*, la podem obtindre a partir de la taula 3, mitjançant una representació dels punts, la determinació d'una línia de tendència (en aquest cas, logarítmica) i l'equació corresponent, com mostra el gràfic 11, on $x = z$ i $y = IDL$.

GRÀFIC 12. REPRESENTACIÓ DELS VALORS D'IDL I Z, CORBA DE TENDÈNCIA I EQUACIÓ LOGARÍTMICA



Amb aquesta aproximació, podríem formular la següent fórmula provisional simplificada:

$$IDL = 44,12 \ln \left[\frac{\log TU}{\log(2^{Tc} - 1)} \right] + 13,084$$

Una proposta que és, naturalment, provisional.

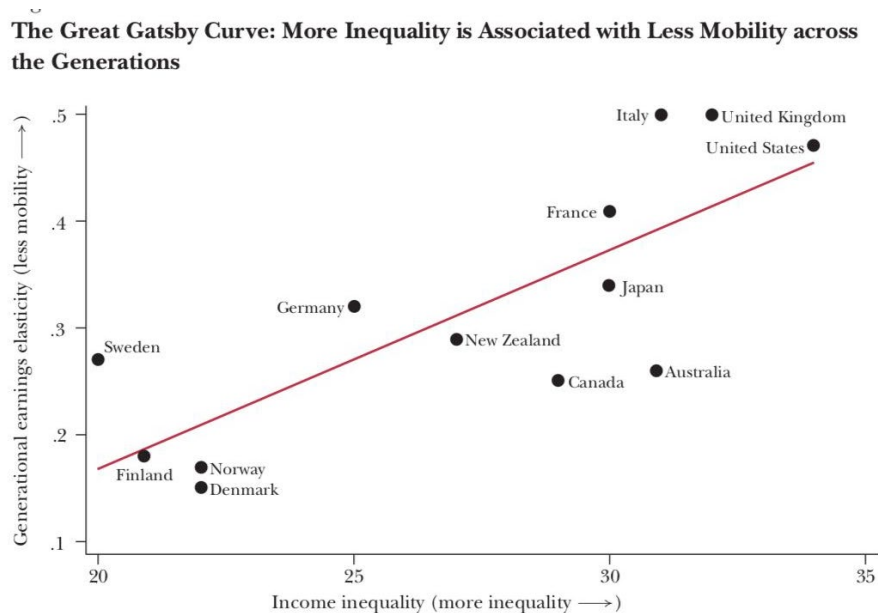
En segon lloc, podem plantejar el debat, relativament recent, sobre l'anomenada Corba del Gran Gatsby, aplicada al cas de la sociolingüística de la nostra llengua.

El celebre economista Paul Krugman qualificà com «very illuminating –and disturbing» aquesta corba, que va descriure així:

A l'eix horitzontal hi ha el coeficient de Gini, una mesura de la desigualtat. A l'eix vertical hi ha l'elasticitat intergeneracional dels ingressos: quant afecta un augment de l'1 per cent dels ingressos del teu pare als teus ingressos esperats; com més gran és aquest número, menor és la mobilitat social. (Krugman 2012)

Ell va fer servir el gràfic 13:

GRÀFIC 13. LA CORBA DEL GRAN GATSBY



Font: Twitter @paulkrugman [27 de febrer de 2017]

En realitat, la corba fou popularitzada per Alan Krueger, en una conferència al Center for American Progress, que havia preparat amb Reid Stevens i Miles Corak. Fou aquest qui proposà el nom, basant-se en el conegut personatge de la novel·la de Scott Fitzgerald, al qual se li feia difícil ascendir de classe social malgrat disposar de diners. Aleshores la corba mostra la relació entre desigualtat econòmica i immobilitat social. El descobriment del grup de Krueger era que les dos variables –desigualtat i immobilitat– presentaven una alta correlació per a un grup de països, amb una línia de tendència recta que presentava una equació $y = 2,2x - 0,27$ i un valor de $R^2 = 0,76$, la qual cosa significa un coeficient de correlació de Pearson⁵ de $R = 0,87$. Ha sigut criticada l'ús polític, poc rigorós, de la corba (Carabaña 2023), encara que considerem que aporta una línia de recerca interessant.

A continuació explicarem quin és el resultat de traslladar la Corba del Gran Gatsby a l'estudi de la llengua. Ja hem explicat com podem traslladar l'índex de Gini a un *IDL*. El problema es planteja amb l'*elasticitat intergeneracional dels ingressos* en termes lingüístics, altrament dit, la dificultat de guanyar més que els progenitors en termes econòmics o, al nostre cas, en termes competencials. En primer lloc, disposem de la *TC* (oral activa) i també de la que anomenarem *Taxa de Transmissió (TT)*, que farem equivalent al percentatge de respostes «amb la família» en la pregunta per l'entorn principal d'aprenentatge, que es pot formular amb la qüestió: «On ha après principalment a parlar el valencià que sap actualment?» de les enquestes esmentades. La diferència entre la *TC* i la *TT* ens indicarà, lògicament, el percentatge de persones competents que han assolit la competència fora de l'entorn familiar, val a dir, han fet un *guany*, més enllà de la transmissió familiar, i podem definir la dificultat (que anomenarem $\mathfrak{X}$) precisament com l'invers d'aquest grup (calculat com la diferència amb la unitat). Podem formular-ho i simplificar-ho així:

$$\mathfrak{X} = 1 - (TC - TT)$$

D'aquesta manera es pot compondre la taula següent.

⁵ El coeficient de correlació de Pearson (R) és un estadístic que expressa la relació entre dues variables. Quan aquesta relació és intensa, s'acosta al valor 1; quan la relació és baixa, s'acosta al valor 0. Si la relació és directa, els valors són positius; si és inversa, adopta valors negatius. El quadrat de R s'anomena coeficient de determinitat.

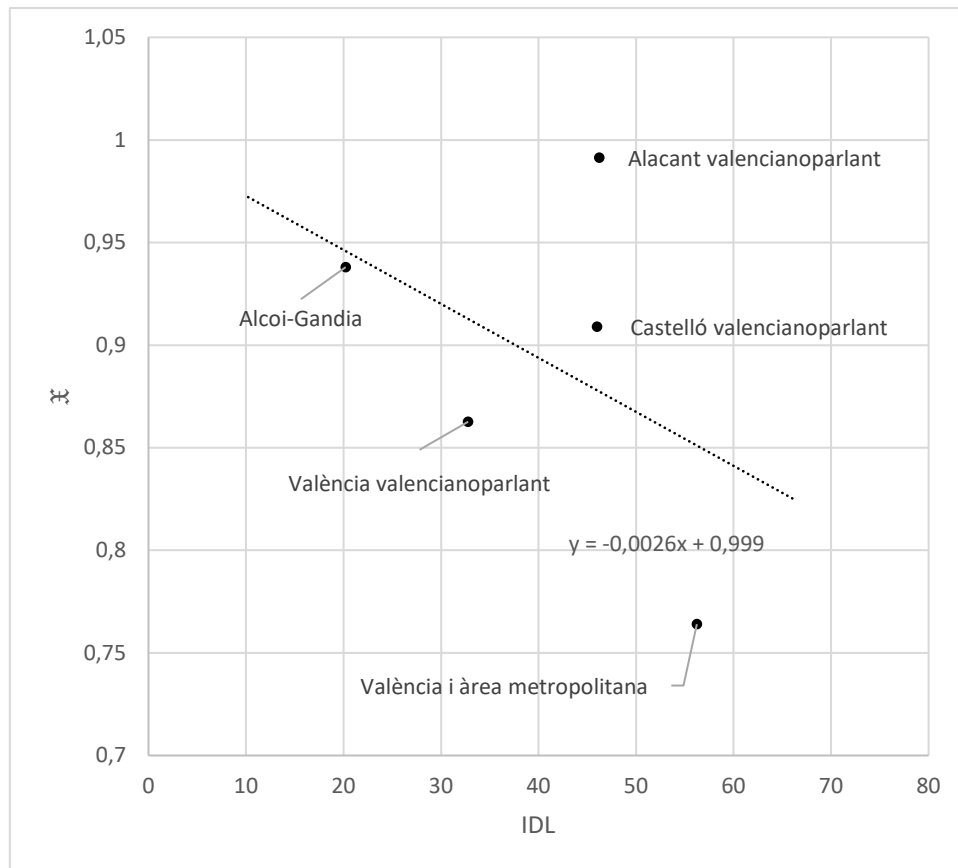
TAULA 6. TC, TU, TT, $\mathfrak{X}$ i IDL, DE LES REGIONS SOCIOLINGÜÍSTIQUES DE LA ZONA
VALENCIANOPARLANT

	<i>TC</i> <i>ponderada</i>	<i>TU</i> <i>ponderada</i>	<i>TT</i>	$\mathfrak{X}$	<i>IDL</i>
Regió d'Alacant (valencianoparlant)	0,37266	0,07716	0,36400	0,99134	46,23176
Regió d'Alcoi-Gandia	0,63800	0,49783	0,57600	0,93800	20,22042
Regió de València i l'àrea metropolitana	0,60900	0,17300	0,37300	0,76400	56,23366
Regió de València (valencianoparlant)	0,68333	0,46166	0,54600	0,86267	32,78304
Regió de Castelló (valencianoparlant)	0,60900	0,26100	0,51800	0,90900	46,01442

Font: Elaboració pròpia d'ECUV 2021.

Amb les dades de $\mathfrak{X}$ i *IDL* podem composar el gràfic 14, on s'ha marcat la línia recta de tendència.

GRÀFIC 14. (CONTRA)CORBA DEL GRAN GATSBY PER AL CAS VALENCIÀ



Font: Elaboració pròpia de la taula 6.

Com s'hi pot veure, estem davant de la que podríem anomenar una *Contracorba del Gran Gatsby*. L'equació de la recta $y = -0,0026x + 0,999$ indica que es tracta d'una recta decreixent, per la qual cosa el coeficient de correlació presenta un valor negatiu, exactament: $R = -0,42819$, el que resulta un coeficient notable per a recerques de ciències humanes i socials.

5.- Pronòstic de dades amb Intel·ligència Artificial

Per a fer un pronòstic de les dades de l'any 2025 de *TC* i *TU*, per tal de calcular els *IDL* corresponents a les diferents regions, utilitzarem un programa d'Intel·ligència Artificial, anomenat Random Forest.

Random Forest és un algorisme d'aprenentatge supervisat, que es presenta com un enfortiment dels *Tree Predictors*. Els *Tree Predictors* són mètodes d'aprenentatge automàtic, que prenen com a base els *Decisions Tree*. Aquests són procediments que divideixen el conjunt de dades en subconjunts més xicotets i, per dir-ho així, aprenen de les característiques dels subconjunts. Amb això poden prendre decisions basades en seqüències *if-then-else*. Per això es fan servir en problemes tant de classificació⁶ com de regressió. Als models *Tree Predictors* s'utilitza un únic arbre de decisió. Per això s'han formulat models més perfeccionats que fan servir uns *Tree Ensembles*, com és el cas de Random Forest, Gradient Boosting i AdaBoost. Aquests *Ensemble Methods* utilitzen tècniques que combinen les prediccions de diversos models base per a millorar la generalització i la precisió⁷.

Random Forest va ser proposat i desenvolupat per Leo Breiman, un estadístic i professor de la Universitat de Califòrnia a Berkeley. Breiman va introduir Random Forest en un article titulat «Random Forests» (Breiman 2001). A continuació es tradueix el resum d'aquest article que acumula a hores d'ara seixanta tres mil citacions.

Els *Random Forests* són una combinació de predictors d'arbres de manera que cada arbre depèn dels valors d'un vector aleatori mostrejat de manera independent i amb la mateixa distribució per a tots els arbres del bosc. L'error de generalització per als boscos convergeix quasi segurament fins a un límit a mesura que el nombre d'arbres al bosc es fa gran. L'error de generalització d'un bosc de classificadors d'arbres depèn de la força dels arbres individuals del bosc i de la correlació entre ells. L'ús d'una selecció aleatòria de funcions per dividir cada node produeix taxes d'error que es comparen favorablement amb Adaboost (Freund & Schapire 1996), però són més robustes, pel que fa al soroll. Les estimacions internes controlen l'error, la força i la correlació i s'utilitzen per mostrar la resposta a l'augment del nombre de funcions utilitzades en la divisió. Les

⁶ Un popular exemple d'arbre de decisió amb objectius classificatoris és allò que s'anomena precisament «triatge» en les entrades d'urgències dels hospitals. Mitjançant una sèrie de preguntes i mesures (temperatura corporal, pressió sanguínia, saturació d'oxigen en sang, etc.) es decideix no només l'especialitat que ha d'atendre el malalt, sinó també el seu codi d'urgència.

⁷ Podríem imaginar una persona que en accedir a les urgències d'un hospital passe per diversos serveis de «triatge», cadascú dels quals hi aplique, a més, una metodologia diferent.

estimacions internes també s'utilitzen per mesurar la importància de la variable. Aquestes idees també són aplicables a la regressió.

La idea darrere de Random Forest era abordar problemes de sobreajustament comuns als arbres de decisió individuals i millorar la precisió i estabilitat del model combinant múltiples arbres entrenats de forma independent. En síntesi, doncs, Random Forest és un algorisme d'aprenentatge supervisat utilitzat per a problemes de classificació i regressió. Pertany a la categoria de mètodes de conjunt, que combina prediccions de múltiples models base per millorar la precisió i la robustesa generals.

Actualment, Random Forest està implementat a diverses biblioteques i plataformes de programació. Aquestes biblioteques i plataformes permeten als desenvolupadors utilitzar l'algorisme en els projectes i anàlisi de dades. Alguns dels entorns i biblioteques més populars amb implementacions de Random Forest són: Scikit-learn (Python), R (R Language), Tensor Flow (Python), H2O.ai, Apache Spark MLlib; Weka (Waikato Environment for Knowledge Analysis). El programa Orange Data Mining 3.36.1, utilitzat per nosaltres, incorpora entre els seus algorismes de modelització Random Forest, així com d'altres esmentats: Tree, AdaBoost, Gradient Boosting, etc.

Per a poder executar *Random Forest*, en el cas de les competències lingüístiques, cal seguir els passos següents:

1r) Preparar un full de càlcul amb les dades de les variables.- Amb les dades disponibles, (ECUV 2005, 2010, 2015 i 2021) s'ha preparat una primera base de dades que ha sigut depurada per tal d'eliminar les respostes NS/NC (uns valors mínims que es distribueixen en cada cas de manera ponderada entre les altres respostes). Es tracta d'una base de dades relativament xicoteta, perquè inclou, a més de la capçalera, les variables esmentades: dues de categorials (competència i regió), la resposta com a variable numèrica ponderada, la data de l'enquesta (l'1 de gener de l'any de referència) i la freqüència de les respostes (en expressió decimal, val a dir, entre 0,0 i 1,0). Cal tindre la precaució de substituir la cometa que indica la part decimal per un punt, segons la tradició anglosaxona. En total, disposem de 321 files ($1 + 4^3 \times 5$) i 6 columnes, val a dir, una matriu de 1.926 dades.

2n) Introduir les dades d'ensinistrament en l'àrea de treball del programa, mitjançant el *widget File*, que pot carregar un full de càlcul depurat, com l'esmentat en el punt anterior.

3r) Verificar la seua correcta determinació (variables categorials, numèriques, com ara el *valor*, i la *data*) per mitjà del *widget Table*.

4t) Transformar les dades amb el *widget Select Columns*. En el cas de competències, totes les variables són *features*, llevat del *valor (target)* i la *data (metas)*.

5é) Enviar les dades transformades al *widget Random Forest*. Aleshores el programa aplica diversos arbres sobre les dades, per tal d'ensinistrar-se sobre la seua constitució.

6é) A continuació s'introdueix una altra base de dades amb la variable data amb el temps futur de la predicció i «?» per al valor a predir.

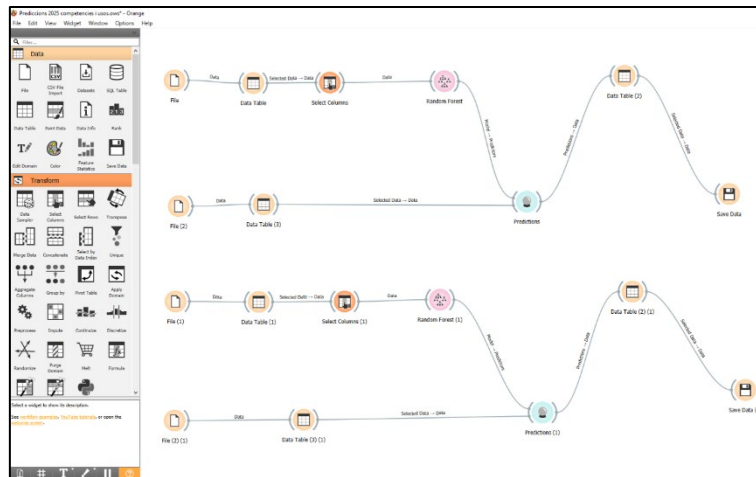
7é) Es vinculen els *widgets* de *Random Forest* i el *widget* d'avaluació *Predictions*. Val a dir, *Predictions* agafa el model que ha determinat *Random Forest* i el calcula per a la segona base de dades, la que correspon a les dades futures.

8é) L'eixida de *Predictions* ha de ser una *Data Table* i un *widget Save Data* a fi de procedir a l'exportació dels resultats. el que produeix un nou fitxer

Per tal de fer una predicció d'ús lingüístic futur, cal repetir les operacions amb les dades d'usos (com s'ha explicat adés canvia la ponderació perquè canvien les categories). En la selecció de columnes, totes les variables són *features*, llevat també del *valor (target)* i la *data (metas)*.

L'aparença de la interfície d'aquest projecte, amb competències (part superior de l'àrea) i usos (part inferior), s'arreplega en la imatge 1:

IMATGE 1. ÀREA DE TREBALL DE RANDOM FOREST



Els resultats de la predicció s'arreglen en la taula 7. Cal recordar que els valors s'han calculat amb la mateixa ponderació emprada adés, on també s'han afegit els valors corresponents per al càlcul de l'IDL.

TAULA 7. IDL PER A LES REGIONS. PRONÒSTIC DEL 2025 AMB IA (MODEL EXPONENCIAL)

	<i>TC ponderada</i>	<i>TU ponderada</i>	<i>z</i>	<i>IDL</i>
Regió d'Alacant (valencianoparlant)				
Pronòstic 2025	0,46575	0,13256	2,09427	46,17000
Regió d'Alcoi-Gandia				
Pronòstic 2025	0,72889	0,52003	1,55866	33,27104
Regió de València i l'àrea metropolitana				
Pronòstic 2025	0,54322	0,18847	2,13239	46,90585
Regió de València (valencianoparlant)				
Pronòstic 2025	0,72118	0,51824	1,51792	32,05119
Regió de Castelló (valencianoparlant)				
Pronòstic 2025	0,63764	0,33634	1,85507	42,82457

Font: Elaboració pròpia

6.- Resultats generals

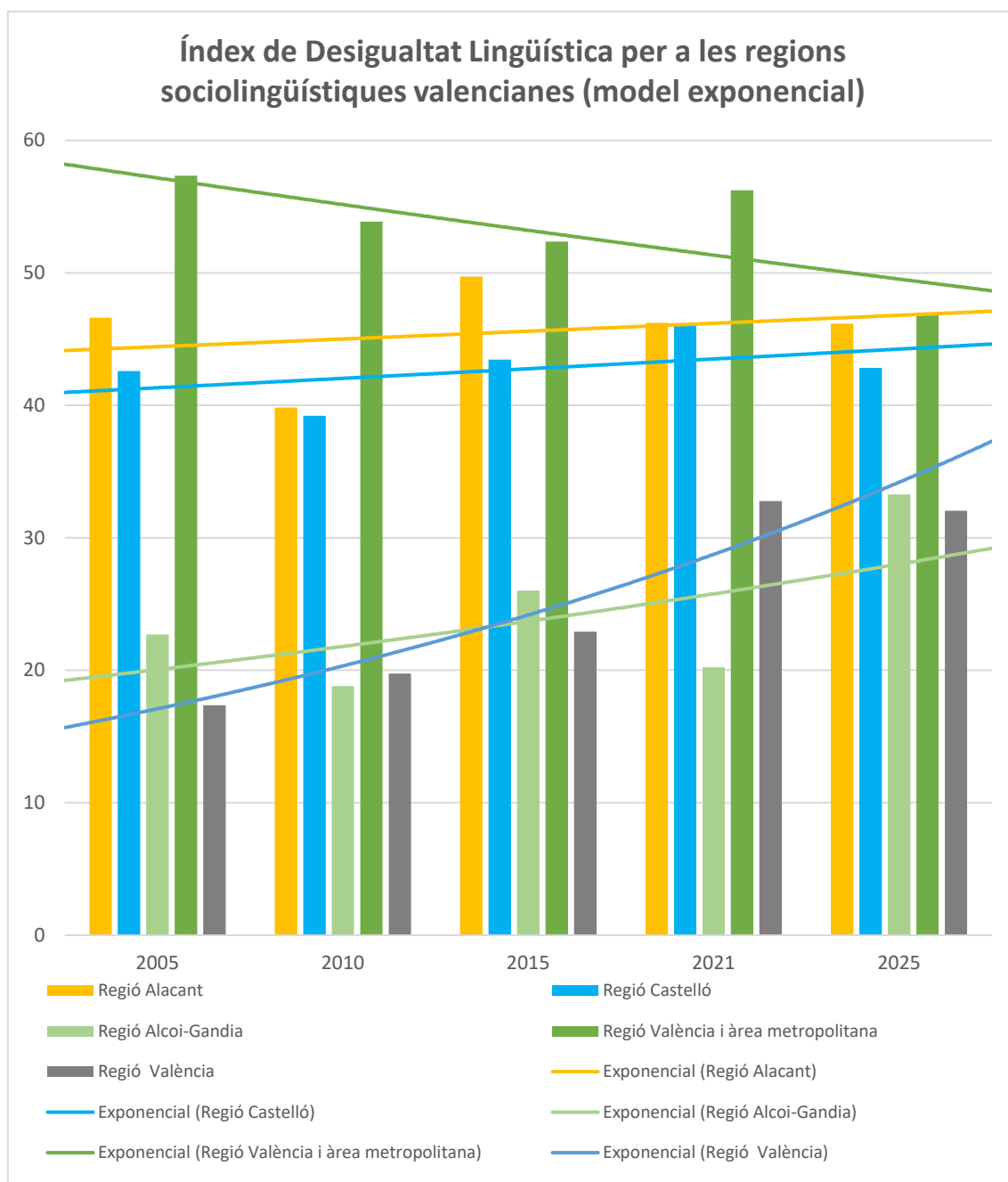
Amb les dades anteriors, podem formar la taula sintètica 8 i el gràfic 15.

TAULA 8. IDL PER A LES REGIONS (MODEL EXPONENCIAL)

	2005	2010	2015	2021	2025*
Regió d'Alacant	46,61	39,83	49,72	46,23	46,17
Regió de Castelló	42,59	39,21	43,44	46,01	42,82
Regió d'Alcoi-Gandia	22,71	18,81	26,02	20,22	33,27
Regió de València i àrea metropolitana	57,34	53,86	52,35	56,23	46,91
Regió de València	17,36	19,75	22,92	32,78	32,05

Font: Síntesi taula 3 i taula 7(*).

GRÀFIC 15. RESULTAT GENERAL (MODEL EXPONENCIAL)



Font: Elaboració pròpia de la taula 8.

En el cas del model de la corba parabòlica, els resultats són:

TAULA 9. IDL PER A LES REGIONS. PRONÒSTIC DEL 2025 AMB IA (MODEL PARABÒLIC)

	<i>TC ponderada</i>	<i>TU ponderada</i>	<i>Paràmetre a</i>	<i>Paràmetre c</i>	<i>IDL</i>
Regió d'Alacant (valencianoparlant)					
Pronòstic 2025	0,46575	0,13256	0,5619	0,4262	52,96421
Regió d'Alcoi-Gandia					
Pronòstic 2025	0,72889	0,52003	0,3149	0,5497	29,67683
Regió de València i l'àrea metropolitana					
Pronòstic 2025	0,54322	0,18847	0,5406	0,4368	50,97017
Regió de València (valencianoparlant)					
Pronòstic 2025	0,72118	0,51824	0,3045	0,5549	28,69631
Regió de Castelló (valencianoparlant)					
Pronòstic 2025	0,63764	0,33634	0,4264	0,4939	40,2033

Font: Elaboració pròpia

El que permet formar la taula de síntesi 10 i el gràfic 16.

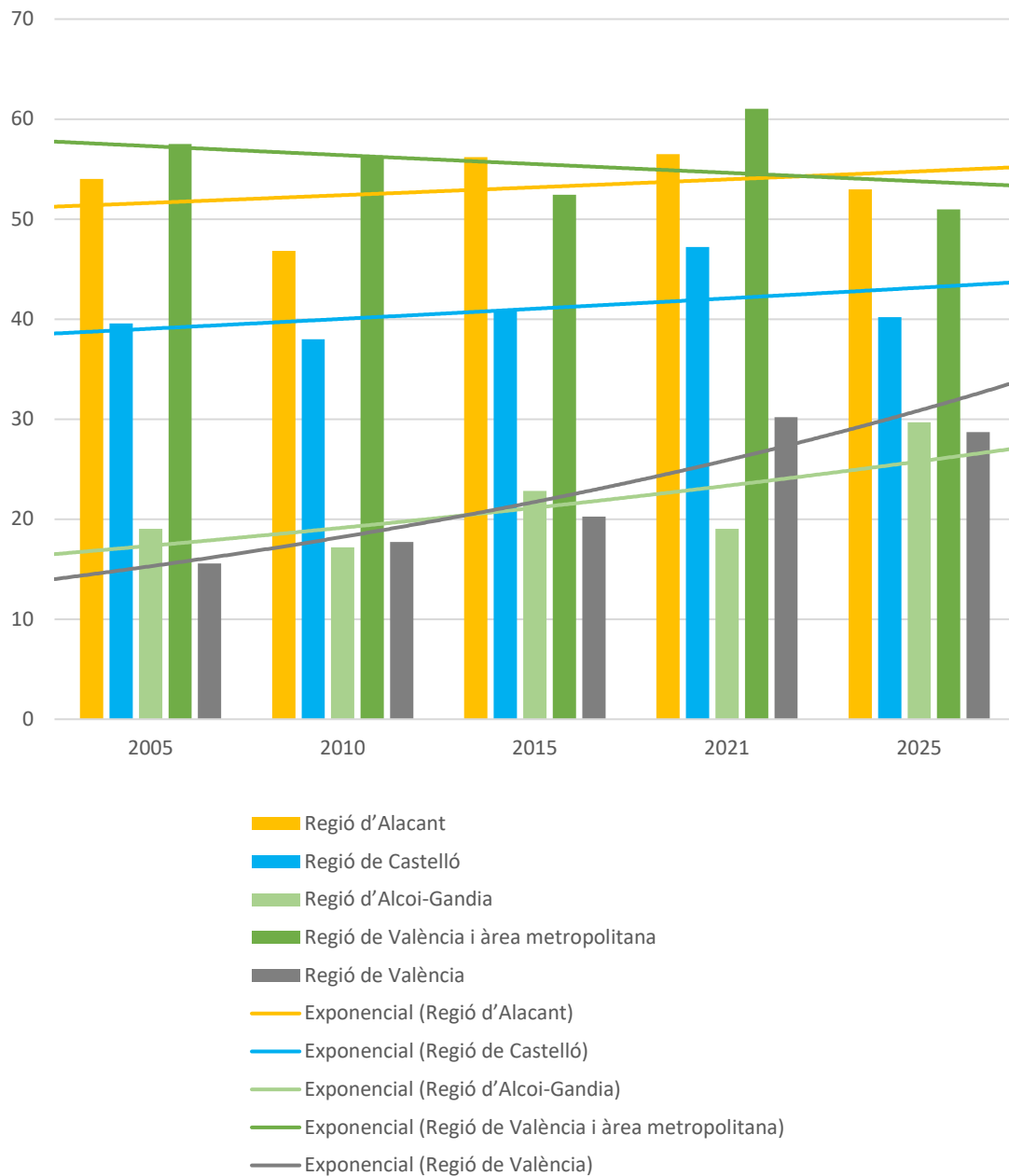
TAULA 10. IDL PER A LES REGIONS I ENQUESTES (MODEL PARABÒLIC)

	2005	2010	2015	2021	2025
Regió d'Alacant	54,01	46,83	56,21	56,49	52,96
Regió de Castelló	39,56	37,99	40,84	47,20	40,20
Regió d'Alcoi-Gandia	19,05	17,18	22,84	19,05	29,68
Regió de València i àrea metropolitana	57,52	56,16	52,43	61,04	50,97
Regió de València	15,57	17,73	20,25	30,19	28,70

Font: Elaboració pròpia

GRÀFIC 16. RESULTAT GENERAL (MODEL PARABÒLIC)

Índex de Desigualtat Lingüística per a les regions sociolingüístiques valencianes (model parabòlic)



Una ulterior comprovació ens permet afirmar que entre els dos models, l'exponencial i el parabòlic, hi ha una alta correlació, amb un **coeficient de Pearson del $R = 0,97428$** .

7.- Bibliografia

- ALEKSANDROV, A. D.; KOLMOGOROV, A. N.; LAVRENTIEV, M. A. (2015): *Le Matematiche*. Torí: Bollati Boringhieri.
- Banc Mundial (2024): Índice de Gini. Disponible en: <https://datos.bancomundial.org/indicador/SI.POV.GINI> (accés: 10 de gener del 2024).
- BENHAMROUCHE, Aziz; MARTÍN VIDE, Javier (2012): «Avances metodológicos en el análisis de la concentración diaria de la precipitación en la España peninsular», *Anales de Geografía de la Universidad Complutense*, 32 (1) 11-27.
- BREIMAN, L. (2001): Random Forests. *Machine Learning*, 45, 5-32. <https://doi.org/10.1023/A:1010933404324>
- CARABAÑA, Julio (2023): Desigualdad y movilidad de ingresos. Ciencia y política de la «curva del Gran Gatsby», *RES. Revista Española de Sociología*, 32, 1.
- ECUV (2005): (Servei d'Investigació i Estudis Sociolingüístics - Conselleria d'Educació, Cultura i Esport) (2005). Enquesta 2005. Sobre coneixement i ús social del valencià (síntesi de resultats) [en línia]. València: Generalitat Valenciana. Conselleria d'Educació, Investiga-ció, Cultura i Esport. <<https://ceice.gva.es/va/web/dgplgm/enquestes-situacio-valencia>>
- (2010): Enquesta 2010. Sobre coneixement i ús social del valencià (síntesi de resultats) [en línia]. València: Generalitat Valenciana. Conselleria d'Educació, Investigació, Cultura i Esport. <<https://ceice.gva.es/va/web/dgplgm/enquestes-situacio-valencia>>
- (2015): Enquesta 2015. Sobre coneixement i ús social del valencià (síntesi de resultats) [en línia]. València: Generalitat Valenciana. Conselleria d'Educació, Investigació, Cultura i Esport. <<https://ceice.gva.es/va/web/dgplgm/enquestes-situacio-valencia>>

- (2021): Enquesta 2021. Sobre coneixement i ús social del valencià (síntesi de resultats) [en premsa]. València: Generalitat Valenciana. Conselleria d'Educació, Investigació, Cultura i Esport.
- EUROSTAT (2023): Income quintile share ratio S80/S20 for disposable income by sex and age group - EU-SILC survey [ilc_di11] (accés 19 octubre 2023).
- FREUND, Y.; SCHAPIRE, R. E. (1996): Experiments with a New Boosting Algorithm. International Conference on Machine Learning, Bari, 3-6 July 1996, 148-156.
- GRAJALES CONESA, Julieta; ACEVES-CHONG, Lorena; RINCÓN-RABANALES, Manuel; CRUZLÓPEZ, Leopoldo (2016): «Jatropha curcas flowers from southern Mexico: chemical profile and morphometrics», *Revista Mexicana de Biodiversidad*, vol. 87, pp. 1321-1327.
- HERNÁNDEZ, Francesc J. (2020): Adaptació del coeficient de GINI per a la formulació d'un coeficient de desigualtat lingüística. *Arxius de sociologia*, núm. 42, 235-263. Disponible en: <<https://roderic.uv.es/handle/10550/76706>>.
- INE (2023): Atlas de distribución de renta de los hogares. Resultados por comunidades autónomas, provincias e islas. Índice de Gini y Distribución de la renta P80/P20. <https://www.ine.es/up/zEkXkJIZ> (accés 19 octubre 2023).
- KRUGMAN, Paul (2012): «The Great Gatsby Curve», *New York Times*, 15/01/2012. Disponible en: <https://archive.nytimes.com/krugman.blogs.nytimes.com/2012/01/15/the-great-gatsby-curve/?referer=&mcubz=3>
- MALDONADO, Antonio; PÉREZ-OCÓN, Rafael; HERRERA, Amparo (2007): «Depression and cognition: New insights from the Lorenz curve and the Gini index», *International Journal of Clinical and Health Psychology*, vol. 7, núm. 1, pp. 21-39.
- NAVARRO, Joaquín (2011): *Al otro lado del espejo. La simetría en matemáticas*. Barcelona: RBA.
- OECD (2023): Income inequality (indicator). doi: 10.1787/459aa7f1-en (accés 20 octubre 2023).

- PÉREZ LABAJOS, Carlos A.; AZOFRA, M.; BLANCO ROJO, Beatriz; ACHÚTEGUI RODRÍGUEZ, Juan José; EGUÍA, E.; DÍAZ, D. (2005): «Collision of fishing vessels. Lorenz curves and GINI indices», *Journal of maritime research*, 2 (3),.
- PIKETTY, Thomas (2013): *Le Capital au XXIe siècle*. París: Éditions du Seuil.
- (2023): *Nature, culture et inégalités*. Nanterre: Société d'ethnologie.
- PONTRIAGUIN, L. S. (2011a): *Cálculo infinitesimal*. Moscou: Krasand.
- (2011b): *Método de coordenadas*. Moscou: Krasand.

8.- Llista de taules, gràfics, requadres i imatges

Taula 1. Coordenades originals i coordenades després de la translació

Taula 2. Coordenades després de la translació i la rotació

Taula 3. IDL per a les regions i enquestes (corba exponencial)

Taula 4. Coordenades després de la translació i la rotació en el cas

Taula 5. IDL per a les regions i enquestes (corba parabòlica)

Taula 6. TC, TU, TT, $\bar{x}$ i IDL de les regions sociolingüístiques de la zona valencianoparlant

Taula 7. IDL per a les regions. Pronòstic del 2025 amb IA (model exponencial)

Taula 8. IDL per a les regions i enquestes (model exponencial)

Taula 9. IDL per a les regions. Pronòstic del 2025 amb IA (model parabòlic)

Taula 10. IDL per a les regions i enquestes (model parabòlic)

Gràfic 1. Quadrat unitat

Gràfic 2. Diagonal creixent

Gràfic 3. Diagonal decreixent

Gràfic 4. Representació del càlcul de l'Índex de Gini

Gràfic 5. Representació del càlcul de l'índex de desigualtat lingüística amb una corba exponencial

Gràfic 6. Model de la corba parabòlica

Gràfic 7. Translació d'eixos cartesianes

Gràfic 8. Rotació d'eixos cartesianes. Representació general

Gràfic 9. Rotació d'eixos cartesianes, representació per al nostre cas

Gràfic 10. Representació de la corba exponencial (Alacant, EUCV 2005)

Gràfic 11. Representació de la corba parabòlica (Alacant, EUCV 2005)

Gràfic 12. Representació dels valors d'IDL i z, corba de tendència i equació logarítmiques

Gràfic 13. La corba del Gran Gatsby

Gràfic 14. (contra)corba del Gran Gatsby per al cas valencià

Gràfic 15. Índex de Desigualtat Lingüística per a les regions sociolingüístiques valencianes (model exponencial)

Gràfic 16. Índex de Desigualtat Lingüística per a les regions sociolingüístiques valencianes (model parabòlic)

Requadre 1. Regions sociolingüístiques de la zona valencianoparlant

Imatge 1. Àrea de treball de Random Forest

Taula

ÍNDIX DE DESIGUALTAT LINGÜÍSTICA DE LES REGIONS SOCIOLINGÜÍSTIQUES VALENCIANES	1
Síntesi	1
ANNEX METODOLÒGIC	5
1.- La mesura de la desigualtat	6
2.- Adaptació de l'Índex de Gini en un Índex de Desigualtat Lingüística	13
3.- Les dades de les regions sociolingüístiques valencianes	22
4.- Ampliació i debat	29
5.- Pronòstic de dades amb Intel·ligència Artificial	34
6.- Resultats generals	38
7.- Bibliografia	43
8.- Llista de taules, gràfics, requadres i imatges	46