Ítaca. Revista de Filologia | ISSN 2386-4753
Núm. 15, 2024, p. 245-267 | https://doi.org/10.14198/itaca.26599

Aplicació de la Intel·ligència Artificial a la recerca sociolingüística del valencià

Application of Artificial Intelligence to Valencian sociolinguistic research

Marc Hernàndez
Ethos Insight, Espanya
insightethos@gmail.com
https://orcid.org/0009-0005-3399-4878
Francesc J. Hernàndez
Universitat de València, Espanya
francesc.j.hernandez@uv.es
https://orcid.org/0000-0001-5229-2998

Citació: Hernàndez, Marc i Hernàndez, Francesc J. (2024). Aplicació de la Intel·ligència Artificial a la recerca sociolingüística del valencià. *Ítaca. Revista de Filologia*, (15), 245-267. https://doi.org/10.14198/itaca.26599

Resum: L'article presenta dues aplicacions de la Intel·ligència Artificial per a la recerca sociolingüística sobre el valencià, una realitzada amb la mineria de dades i una altra amb xarxes neuronals. S'expliquen aquestes dues funcions de la Intel·ligència Artificial i es detalla com s'ha procedit en els dos casos. Aquestes aplicacions es realitzen amb les dades depurades de les Enquestes de Coneixement i Ús del Valencià (2005, 2010, 2015 i 2021) i amb recerques pròpies fetes amb enquestes i treball de camp, i es refereixen a prediccions, expectatives de competència aliena i indicadors unitaris de competència i ús, així com índex de desigualtat lingüística. Per a la mineria de dades s'han calculat les prediccions per a 2025, de competències i usos lingüístics, el que permet fer un pronòstic de l'evolució d'un índex de desigualtat lingüística referit a les diverses regions sociolingüístiques valencianes. Aquest es fonamenta en relació al conegut Índex de Gini de desigualtat socioeconòmica, que, en el nostre cas, es calcula com a Índex de Desigaltat Lingüística amb un model de corba exponencial. S'hi aporta la fonamentació matemàtica del càlcul. En aquest punt destaquem el problema de la infravaloració de les competències lingüístiques alienes, un factor del decalatge entre competència i ús. A continuació fem una aplicació de xarxes neuronals, basades en un model d'aprenentatge. S'explica

el fonament de les xarxes neuronals, el procediment per al seu ensinistrament i l'aplicació a les dades de competències i usos lingüístics de les regions sociolingüístiques del valencià. S'hi aporten les dades resultants i es presenta la interessant la correlació final entre aquell Índex de Desigualtat Lingüística i les simulacions de xarxes neuronals segons el model d'aprenentatge especificat. Aquesta conclusió és summament important, perquè permet simplificar la determinació de la desigualtat lingüística com a un càlcul de simulació de xarxes neuronals, el que obri perspectives innovadores a la recerca lingüística.

Paraules clau: sociolingüística valenciana; intel·ligència artificial; mineria de dades; xarxes neuronals; índex de desigualtat lingüística; valencià; català.

Abstract: The article presents two applications of Artificial Intelligence for sociolinguistic research on Valencian, one carried out with data mining and the other with neural networks. These two functions of Artificial Intelligence are explained and the procedure used in both cases is detailed. These applications are carried out with the cleaned data from the Enquestes de Coneixement i ús del Valencià (2005, 2010, 2015 and 2021) and with own research carried out with surveys and fieldwork, and refer to predictions, expectations of foreign competence and unit indicators of competence and use, as well as an index of linguistic inequality. For data mining, predictions have been calculated for 2025, for language competences and uses, which allows a forecast to be made of the evolution of an index of linguistic inequality referring to the different Valencian sociolinguistic regions. This is based on the so-called Gini Index of socio-economic inequality, which, in our case, is calculated as the Index of Linguistic Inequality using an exponential curve model. The mathematical basis of the calculation is provided. At this point we highlight the problem of the underestimation of the value of foreign language skills, a factor in the mismatch between competence and use. We then apply neural networks, based on a learning model. It explains the basis of neural networks, the procedure for their analysis and their application to the data on linguistic competences and uses in the sociolinguistic regions of Valencian. The resulting data are provided and the interesting final correlation between the Index of Linguistic Inequality and the simulations of neural networks according to the specified learning model is presented. This conclusion is extremely important, because it allows us to simplify the determination of linguistic inequality as a neural network simulation calculation, which opens up innovative perspectives in linguistic research.

Keywords: valencian sociolinguistics, artificial intelligence, data mining, neural networks, linguistic inequality index, valencian language, catalan language.

Rebut: 19/12/2023, **Acceptat:** 01/03/2024

1. Introducció

Són moltes les aportacions de la Intel·ligència Artificial (IA) a la recerca sobre el valencià. Resulta sorprenent que la mateixa IA ens puga indicar *automàticament* quines són els possibles usos. Per exemple, li hem preguntat (més exactament, n'hem escrit un *prompt*) al *xatbot* ChatGPT sobre l'ús de les xarxes neuronals en llengües menys esteses, i ens ha proporcionat la resposta que copiem traduïda a continuació:

> Es poden utilitzar les xarxes neuronals en diverses vies a fi d'analitzar l'evolució de llengua. Heus ací algunes aproximacions i idees sobre com podeu emprar xarxes neuronals per a aquesta recerca:
>
> -Anàlisi de Text i Processament de Llenguatge Natural (NLP). Per exemple, anàlisi de sentiment: podeu utilitzar models NLP per a analitzar sentiment en textos de manera diacrònica, la qual cosa pot proporcionar idees sobre els canvis d'actituds cap a la llengua de minoria.
> -Modelatge lingüístic: Ensinistrament de models lingüístics per a generar textos en llengües de minoria i anàlisi diacrònic de la llengua.
> -Reconeixement de noms d'entitats: identificació i seguiment de l'ús d'antropònims, topònims o esdeveniments narrats en la llengua de minoria en textos històrics.
> -Seguiment de mots *embeddings* (amb programes com ara Word2Vec o FastText), per a crear representacions de mots en la llengua de minoria, plasmant l'evolució diacrònica.
> -Classificació de llengua: ús de xarxes neuronals per classificar textos o documents a períodes de temps o generacions diferents per entendre com la llengua ha evolucionat amb el temps.
> -Anàlisi de Mitjans de comunicació socials: recull i anàlisi de dades de mitjans de comunicació socials en la llengua de la minoria per controlar ús de llengua i canvia en l'era digital.
> -Ús xarxes neuronals recurrents (RNN) o xarxes neuronals *convolucionals* (CNN) per a analitzar l'ús de llengua en textos com a sèrie de temps, identificant tendències i patrons.
> -Predicció d'evolució de la llengua: ensinistrament de xarxes neuronals o models de transformació de la llengua, amb dades històriques, per a predir com evolucionarà la llengua en el futur.

> -Estudis creuats: comparació de l'evolució de la llengua de minoria amb altres llengües, tant de minoria com de majoria, per identificar patrons compartits i característiques úniques.
> -Visualització de dada: ús xarxes neuronals per crear visualitzacions aquella evolució de la llengua en la pantalla, fent les troballes més accessibles i comprometent una audiència més ampla.

Deixarem de banda recerques estrictament lingüístiques, com ara les referides a PNL (sobre això, vegeu Torrijos i Sánchez 2023), i ens centrarem en recerques de sociologia de la llengua.

Pel que fa a la recerca sociolingüística, els intents d'aconseguir una modelització satisfactòria (Álvarez 1984; Querol 2002; Fabà 2003; Hernàndez 2015) han sigut escassos i, en tot cas, no es poden considerar definitius. Per això, en aquest article avançarem en les aportacions que hi puga fer la Intel·ligència Artificial. Més concretament, presentarem ací dues recerques relacionades amb allò que s'anomena «mineria de dades» i una altra relacionada amb «xarxes neuronals», en ambdós casos, sobre les dades de les regions sociolingüístiques valencianes. A la fi de l'article, les dues recerques es relacionen satisfactòriament. Ara bé, el camp de la Intel·ligència Artificial és summament dinàmic (Rodríguez 2018) i som conscients que els programes que emprem en aquesta recerca aviat tindran versions més avançades o n'apareixeran noves aplicacions.

No hi ha treballs anteriors en aquesta línia perquè l'ús de xarxes neuronals s'ha emprat en el camp de la sociolingüística per a lingüística de la parla o sociofonètica, que no és el nostre interés, bé la detecció de comunitats lingüístiques —espais semàntics— o per a analitzar la relació amb la variació lingüística. A tall d'exemple hi ha els treballs de Gupta i DiPadova (2019), Kretzschmar (2008), Thomas (2011) i Tsoukala et al. (2020). En tot cas, en el nostre domini lingüístic s'ha fet ús de les xarxes neuronals per a processar quantitats massives de textos i trobar-hi «sentiments» (Balaguer et al. 2019), com ara la «identificació» (Simões et al. 2014). Aquest treball correspon més aviat, però, a la psicolingüística o la psicologia social del llenguatge.

Amb tot i això, els objectius que ens plantegem amb aquest treball són, d'una banda, fer prediccions sociolingüístiques amb «mineria de dades» per al cas del valencià i mostrar-ne la utilitat general. Com s'hi pot veure, amb aquestes dades es pot calcular un índex de desigualtat lingüística, no no-

més amb les dades precedents de les diverses enquestes, sinó també fer-ne un pronòstic futur de la seua evolució. Però a més, l'article té la pretensió d'emprar les xarxes neuronals per a una aproximació a la modelització comentada adés. Aquests dos procediments —mineria de dades i xarxes neuronals— s'integren en el resultat final de l'article, en què podem comprovar l'alta correlació entre els diferents índexs de desigualtat lingüística de les regions sociolingüístiques valencianes i la simulació realitzada amb xarxes neuronals.

2. Una aplicació de la mineria de dades a la sociolingüística del valencià

2.1 Mineria de dades i recerca lingüística

La «mineria de dades» és una metàfora escaient perquè, en definitiva, es tracta d'aconseguir noves dades o veure què podem traure del gran conjunt de dades disponibles. En el cas, però, de la sociolingüística valenciana, les dades són certament escasses. Dissortadament, la sèrie d'enquestes sociolingüístiques no és molt ampla i en molts casos hem perdut les bases de dades originals i només disposem dels informes divulgatius. Tot això aconsella restringir l'aplicació a les dades disponibles de les Enquestes de Coneixement i Ús del Valencià realitzades els anys 2005, 2010, 2015 i 2021 —aquesta lleugerament endarrerida per la pandèmia de la Covid-19— (ECUV 2005, 2010, 2015 i 2021). Farem servir ací les dades sobre competències lingüístiques orals i escrites, actives i passives, en total, es tracta de quatre possibilitats: COP, COA, CEP i CEA, equivalents a *entendre*, *parlar*, *llegir* i *escriure*. El camp d'estudi són les regions sociolingüístiques valencianoparlants: regió de la província de Castelló valencianoparlant, regió de la província de València valencianoparlant, ciutat de València i la seua àrea metropolitana; regió d'Alcoi-Gandia i regió de la província d'Alacant valencianoparlant. Les respostes possibles a les enquestes esmentades són si la persona disposa de la competència *perfectament*, *bastant bé*, *un poc* o *gens*. A continuació explicarem com podem fer prediccions de dades futures amb un programa de «mineria de dades», o més exactament amb l'aplicació anomenada *Random Forest*.

2.2 L'algorisme 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 arbres de decisió, *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 la qual cosa poden prendre decisions basades en seqüències *if-then-else*. Per això s'empren en problemes tant de classificació com de regressió. En els models *tree predictors* es fa servir un únic arbre de decisió. Per això s'han formulat models que fan servir no un arbre de decisions, sinó un *conjunt d'arbres —tree ensembles—*, com és el cas de *Random Forest*, *Gradient Boosting* i *AdaBoost*. Aquests mètodes de conjunt *—ensemble methods—* fan servir tècniques que combinen les prediccions de diversos models base per a millorar la generalització i la precisió. *Random Forest* va ser proposat i desenvolupat per Leo Breiman (Breiman 2001), un estadístic i professor de la Universitat de Califòrnia a Berkeley. A continuació es tradueix el resum d'aquest article que a hores d'ara acumula seixanta tres mil citacions:

> Els boscos aleatoris són una combinació de predictors d'arbres de manera que cada arbre depén 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 depén 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 estimacions internes també s'utilitzen per mesurar la importància de la variable. Aquestes idees també són aplicables a la regressió. (Breiman 2001)

Altrament dit: la idea que hi ha al darrere de *Random Forest* era abordar problemes de sobreajustament comuns als arbres de decisió individuals i

millorar la precisió i estabilitat del model, tot combinant múltiples arbres entrenats de forma independent. En síntesi, *Random Forest* és un algorisme que 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 en diverses biblioteques i plataformes de programació. Aquestes biblioteques i plataformes permeten que els desenvolupadors puguen utilitzar l'algorisme en els seus projectes i en les anàlisis de dades. Alguns dels entorns i biblioteques populars on hi ha implementacions de *Random Forest* inclouen: Scikit-learn (Python), R (R Language), Tensor Flow (Python), H2O.ai, Apache Spark MLlib, Weka (Waikato Environment for Knowledge Analysis).

2.3 Utilització en la recerca sociolingüística

El programa Orange Data Mining 3.36.1, que hem fet servir ací, incorpora entre els seus algorismes de modelització *Random Forest*, així com d'altres esmentats adés: *Tree*, *AdaBoost*, *Gradient Boosting*, etc., el que permet també una comparació de resultats.

Per a poder executar *Randow 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) hem preparat una primera base de dades que hem depurat per tal d'eliminar les respostes NS/NC (uns valors mínims que es distribueixen en cada cas de manera ponderada entre les altres respostes[1]). El resultat és el fitxer *basedades01.xlsx*. 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 (hem assignat els

1. No es pot fer aquesta distribució amb una ponderació calculada amb la base de dades perquè, en arrodonir les quantitats dels percentatges poden no sumar la unitat. És recomanable fer una distribució intuïtiva del que, d'altra banda, són quantitats mínimes, que no apareixen en totes les Enquestes.

valors: 3/3 a *perfectament*, 2/3 a *bastant bé*, 1/3 a *un poc* i 0/3 a *gens*[2]), 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 i 6 columnes.

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.

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*.

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. El resultat és el fitxer *basedades02.xlsx*.

7é) Es vinculen els *widgets* de *Random Forest* i el *widget* d'avaluació *Predictions*.

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. Llavors es produeix el fitxer *basedades03.xlsx*.

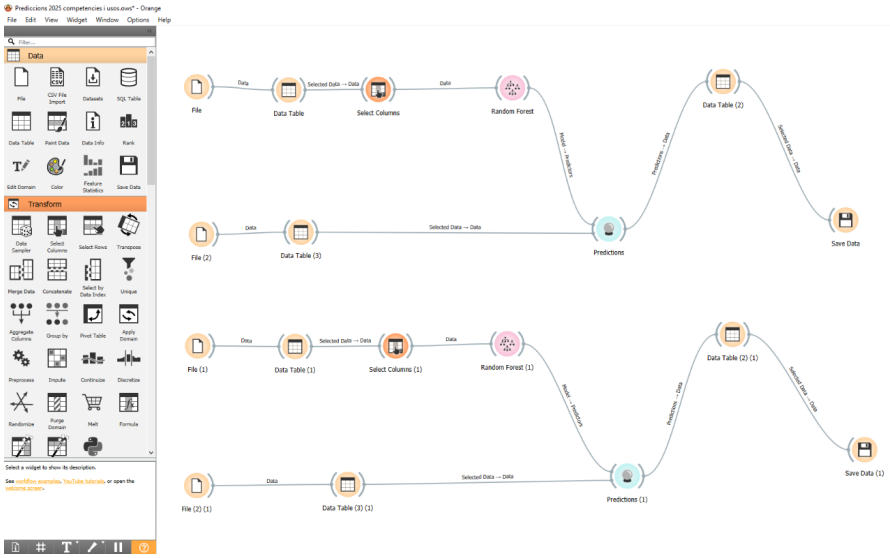
Per tal de fer una predicció d'ús lingüístic futur, cal repetir les operacions. En la selecció de columnes, totes les variables són *features*, llevat també del *valor* (*target*) i la *data* (*metas*).

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. Els resultats de la predicció s'arrepleguen en la taula 1. Cal recordar que els valors s'han calculat amb la mateixa ponderació emprada adés. Si es manté la sèrie d'enquestes, el 2025 comprovarem l'exactitud de la predicció i, en tot cas,

2. Les ponderacions suposen una certa «interpretació» de les respostes possibles. També podria emprar-se una altra sèrie de valors: 4/4 a *perfectament*, 3/4 a *bastant bé*, 1/4 a *un poc* i 0/4 a *gens*, per "obrir" més l'espai central. Amb tot, les diferències en els resultats serien mínimes.

podrem ensinistrar millor l'algorisme amb noves dades perquè les seues prediccions siguen més ajustades.

Imatge 1



Taula 1. Prediccions amb Random Forest

Regió	Competència	Valor
Província d'Alacant valencianoparlant	COP	62,81%
	COA	46,58%
	CEP	51,47%
	CEA	30,60%
Província de Castelló valencianoparlant	COP	79,73%
	COA	63,76%
	CEP	61,55%
	CEA	44,08%
Regió Alcoi-Gandia	COP	81,67%
	COA	72,89%
	CEP	67,23%
	CEA	51,35%

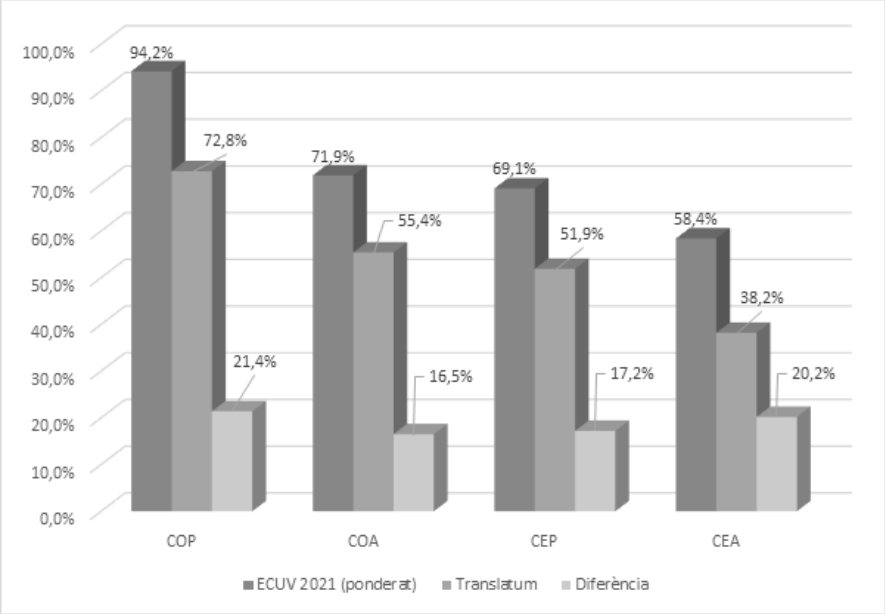
Ciutat de València i àrea metropolitana	COP	74,49%
	COA	54,32%
	CEP	62,76%
	CEA	34,27%
Província de València valencianoparlant	COP	83,74%
	COA	72,12%
	CEP	70,35%
	CEA	53,65%

Font: Elaboració pròpia. COP: Competència Oral Passiva (entendre); COA: Competència Oral Activa (parlar); CEP: Competència Escrita Passiva (llegir); CEA: Competència Escrita Activa (escriure).

3. Un corol·lari sobre un greu problema del valencià

L'any 2022 realitzàrem un estudi sobre la percepció de les competències lingüístiques en valencià de la població, circumscrit a les regions sociolingüístiques de la província de València i la ciutat de València i la seua àrea metropolitana. Aquest projecte de recerca, que anomenàrem *Translatum* (Hernàndez i Hernàndez 2022), aportà una conclusió general interessantíssima: la població valenciana infravalora les competències alienes (el que podem dir: *expectativa de taxes de competència*: *TCexp*). És a dir, que entre les competències mesurades per l'Enquesta de Coneixement i Ús del Valencià del 2021 (les *taxes de competència*: *TC*), ponderades per a les regions de la recerca, i la suposició que la població té d'aquestes competències, es dedueix que els parlants, en considerar que els interlocutors no disposen de competència, fan servir el castellà (i la presència social d'aquesta llengua, al temps, reforçaria la suposició de competències menors). Així es pot apreciar al gràfic 1.

Gràfic 1



Font: Projecte Translatum. COP: Competència Oral Passiva (entendre); COA: Competència Oral Activa (parlar); CEP: Competència Escrita Passiva (llegir); CEA: Competència Escrita Activa (escriure).

En la recerca, partírem de la hipòtesi per a la Competència Oral Activa (*parlar*):[3]

$$TC^2 = TCexp$$

I els resultats de la recerca proporcionà exponents de la entre 1,95 (per a tots els resultats de l'enquesta *Translatum* N=800) i 1,75 (si ponderem la mostra per les dues regions sociolingüístiques estudiades), valors molt aproximats al 2,0 de la hipòtesi. D'aquesta manera, també podem fer una predicció de la per al 2025 que s'arreplega en la taula 2.

3. Sobre la justificació d'aquesta hipòtesi, vegeu Hernàndez 2022.

Taula 2.

Regió	*Valor TC(OA) exp*
Regió de la província d'Alacant valencianoparlant	21,69%
Regió de la província de Castelló valencianoparlant	40,66%
Regió Alcoi-Gandia	53,13%
Regió de la ciutat de València i àrea metropolitana	29,51%
Regió de la província de València valencianoparlant	52,01%

Font: Elaboració pròpia

4. Una aplicació de xarxes neuronals a la sociolingüística del valencià. Validació d'un índex de desigualtat lingüística

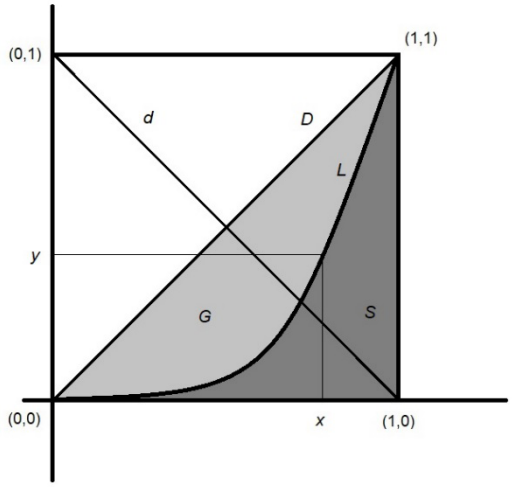
En un article anterior avançàrem en un model probabilístic de relació entre taxes de competència i taxes d'ús per al cas del valencià (Hernàndez 2020a). L'interés de relacionar les dues variables es troba en la possibilitat d'elaborar indicadors de desigualtat lingüística, equiparables als indicadors de desigualtat socioeconòmica acceptats, com ara l'índex de Gini (Hernàndez 2020b). Això exigeix no només pensar la relació entre competències i ús *de manera no lineal*, sinó també fer servir indicadors unitaris, que puguen definir bones correlacions amb totes les variables, com ara el que proposàrem basat en l'equació de l'entropia de Boltzmann (Hernàndez 2023). Doncs bé, aquestes qüestions d'allò més importants per a la recerca sociolingüística de la nostra llengua, es poden mamprendre amb la IA, més concretament amb les simulacions de les xarxes neuronals.

Explicarem en primer lloc, de manera succinta, el càlcul d'un índex de desigualtat lingüística, i després la seua validació amb xarxes neuronals.

4.1 Índex de desigualtat lingüística

Amb les dades resultants (tant les depurades com les estimades), s'aplica la metodologia del càlcul de desigualtat de l'índex de Gini (*IG*), a partir del Coeficient de Gini (*CG*), amb un model de corba exponencial, de manera que, si partim del gràfic 2,

Gràfic 2



Aleshores:

$$IG = 100\ CG$$

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

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

A continuació s'explica el càlcul de *S*. Per a la corba exponencial es considera l'equació més simple:

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

Que acompleix els requisits de la corba *L*: si $x = 0 \rightarrow$ i si $y = 1 \rightarrow x = 1$.

Per a cada parella de valors (x, y), és a dir (TC_i, TU_i), es calcula el corresponent valor de *z*:

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

Amb el valor de *z* calculat, es donen 101 valors (de 0,00 a 1,00) a *x* en l'equació $y = (2^x - 1)^z$. Els valors resultants es representen en uns eixos cartesians i es demana al programa Excel (del paquet Microsoft 365) l'equació de la línia de tendència dels punts representats de tipus polinòmic d'orde 6. Amb aquesta equació, de la forma: $y = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$, es calcula la integral definida (val a dir, la superfície baix de la corba) de la manera simplificada que s'indica a continuació:

$$S = \int_0^1 (ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g)\, dx =$$

$$= a\int_0^1 x^6\,dx + b\int_0^1 x^5\,dx + c\int_0^1 x^4\,dx + d\int_0^1 x^3\,dx + e\int_0^1 x^2\,dx + f\int_0^1 x\,dx + g\int_0^1 dx =$$

$$= a\left[\frac{x^7}{7}\right]_0^1 + b\left[\frac{x^6}{6}\right]_0^1 + c\left[\frac{x^5}{5}\right]_0^1 + d\left[\frac{x^4}{4}\right]_0^1 + e\left[\frac{x^3}{3}\right]_0^1 + f\left[\frac{x^2}{2}\right]_0^1 + g\,[x]$$

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

Amb el valor de *S* es calcula *IG* en cada cas, com s'ha explicat abans. Per a la taula anterior, el procés s'ha repetit 25 vegades, una per a cada cel·la. Els resultats s'arrepleguen en la taula 3.

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

Font: Elaboració pròpia. (*) Estimacions, segons la metodologia indicada.

4.2 Validació amb xarxes neuronals

Farem una explicació introductòria (adaptada de Yoshimi et al. 2023) sobre què és una xarxa neuronal artificial. Una xarxa neuronal artificial (XNA, per *Artificial Neural Network*, ANN) és un model informàtic que presenta aspectes en comú amb una xarxa neuronal biològica (XNB, per *Biological Neural Network*, BNN). Una XNA té nodes (*nodes*) –també: unitats (*units*)– i pesos (*weights*) que són anàlegs a les neurones i les sinapsis de la XNB.

L'activació d'una neurona seria l'equivalent a la freqüència del tret (*firing rate*) o el potencial de membrana (*membrane potential*) d'una neurona real. Els pesos tenen una força. Quan l'activació flueix per una xarxa, els pesos amb força positiva tendeixen a augmentar l'activació; els pesos amb força negativa tendeixen a reduir l'activació. Corresponen a sinapsis excitatòries o inhibitòries (*excitatory and inhibitory synapses*), altrament dit, l'eficàcia sinàptica.

Les activacions dels nodus canvien d'acord amb regles d'activació (*activation rules*) i regles d'aprenentatge (*learning rules*). En alguns models, un node també pot produir una espiga (*spike*). Nodes i pesos formen una estructura de xarxa o matemàtica, també anomenada graf (*graph*). L'estructura de graf formada pels nodes i els pesos d'una xarxa és la topologia (*topology*) de la xarxa. Indicarem les diferències entre la computació habitual i les xarxes neuronals.

En un ordinador clàssic, els símbols discrets (cadenes de 0 i 1, o bits) funcionen de forma seqüencial mitjançant regles. Els bits d'informació es col·loquen als registres de la unitat central de processament (CPU) i s'hi apliquen les regles lògiques del conjunt d'instruccions de la CPU. Els ordinadors es programen a mà per a fer coses útils. L'interior d'una CPU i els sistemes de memòria d'un ordinador són entorns acuradament controlats. No els van bé els senyals sorollosos ni els danys.

La computació en una xarxa neuronal és diferent. La computació neuronal no es fonamenta en operacions seqüencials i basades en regles sobre bits, sinó en operacions paral·leles en què els patrons d'activació dels nodes es transformen mitjançant forces de pes. Les xarxes neuronals també toleren millor els senyals sorollosos i els danys que els ordinadors digitals.

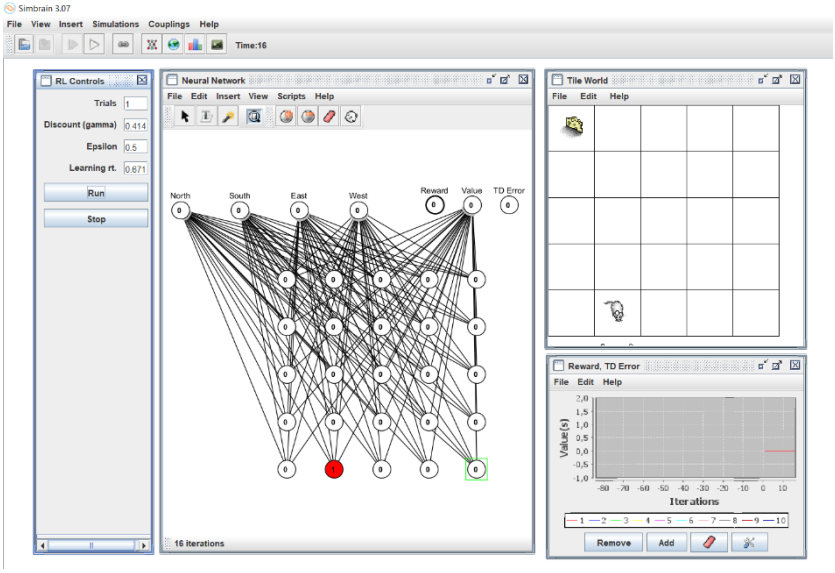
Les xarxes s'entrenen mitjançant aprenentatge, no es programen. Mostrem a la xarxa el que volem que faça i aquesta aprén a fer-ho. Això s'anomena aprenentatge supervisat, ja que coneixem l'eixida correcta per a cada entrada i podem dir a la xarxa exactament quina eixida ha de produir per a qualsevol entrada donada. El positiu d'aquest procediment és que una vegada la xarxa està ensinistrada amb algunes dades és possible la generalització, és a dir, la xarxa pot tractar amb dades noves a les quals mai abans no ha estat exposada.

Les xarxes també poden aprendre sense cap mena de senyal d'ensinistrament o reforç, simplement captant l'estructura estadística de l'entorn (aprenentatge no supervisat). D'altres diferències són: les xarxes neuronals fan un processament en paral·lel; quan pateixen danys, experimenten una degradació gradual (altrament dit: tenen tolerància a fallades); les xarxes tenen representacions distribuïdes en lloc de representacions localistes.

A continuació procedirem a la nostra aplicació (vegeu Tosi i Yoshimi 2016). Podem considerar que les anomenades interaccions lambda –per exemple, una persona amb competència oral activa interacciona amb una que sap parlar valencià– són un cas de l'aprenentatge reforçat, la predicció informàtica del qual fou definida per Sutton el 1996 (vegeu Sutton i Barto, 2015 i sobre el *deep-learning*, Piloto 2022). Ací farem servir un implementació de Jeff Yoshimi i Jonathon Vickrey, disponible en el programa Simbrain 3.7, gràcies al qual la xarxa neuronal pot simular el comportament d'un individu, interpretat segons el model d'un «actor crític». En primer lloc, farem una explicació senzilla del model i, en segon lloc, mostrarem l'aplicació per a les regions sociolingüístiques valencianes.

Imaginem un ratolí virtual ubicat en el cantó d'un quadrat imaginari, dividit en caselles a la manera d'un tauler d'escacs. Suposem que, al cantó oposat, hi ha un tros de formatge. El ratolí, guiat per la *seua* xarxa neuronal virtual (la que hem definit amb els paràmetres), es pot desplaçar a qualsevol de les caselles que té a cada banda (però no en diagonal). De manera aleatòria passarà d'una casella a un altra fins que trobe la del formatge virtual. A continuació, com que hem programat un bon nombre d'intents, el ratolí comença de nou, però ja ha aprés alguna cosa —millor dit: les xarxes neuronals han aprés— i li costarà menys arribar al formatge, però encara hi haurà desplaçaments erronis. Encara que aquest tipus de presentacions o entorns aviat seran obsolets, en la imatge 3 se n'ofereix una visualització:

Imatge 3



En aquest model simple, a més del nombre d'intents, cal definir una sèrie de paràmetres en la xarxa neuronal:

- Factor gamma o factor de descompte: l'orientació al futur de l'agent. El rang és 0-1. Per a valors més pròxims a 0, l'agent està enfocat a recompenses més immediates; per a valors més pròxims a 1, l'agent se centra més en recompenses no immediates.

- Paràmetre èpsilon –cal no confondre'l amb l'indicador èpsilon definit en Hernàndez (2023): la probabilitat que té el ratolí virtual de fer una acció aleatòria. Lògicament aquesta probabilitat oscil·la entre 0 (no té probabilitat de fer accions aleatòries) i 1 (totes les accions en són aleatòries).

- Taxa d'aprenentatge: la proporció de saber (els «pesos», s'hi diu metafòricament) que s'actualitza en cada pas de temps (també oscil·la entre 0 i 1).[4]

4. En el cas del ratolí real, aquests tres paràmetres, el factor gamma, èpsilon i la taxa d'aprenentatge es relacionen amb neurotransmissors, com ara la serotonina, la noradrenalina (també coneguda com a norepinefrina) i l'acetilcolina, que al temps estan relacionats amb tipus d'aprenentatges.

No és difícil fer una adaptació d'aquests paràmetres per intentar una simulació per a les regions sociolingüístiques.

- Factor gamma: que es fa equivalent a la taxa d'ús l'àmbit de la llar (calculat amb la metodologia indicada adés de *Random Forests*).
- Paràmetre èpsilon: Considerarem aquest paràmetre com l'invers de la taxa d'ús al carrer (1 – TUcarrer), calculat de la mateixa manera.
- Taxa d'aprenentatge: la farem equivalent a la taxa de competència, de la competència oral activa (parlar), que també es calcula amb les prediccions de la mineria de dades.

En fer equivalents els paràmetres de la xarxa amb les variables sociolingüístiques podem procedir a l'experimentació. És a dir, demanem al nostre ratolí virtual que realitze una sèrie d'intents i controlem el temps de realització.

Per a explicar quin és el significat d'aquesta experimentació hem de tindre en compte alguns factors. Anteriorment hem definit la desigualtat en termes d'una relació no lineal —una corba exponencial— entre competència i ús, és a dir, la coacció que experimenta un individu competent per poder fer-ne ús. Naturalment, a més coacció, més temps per a aconseguir una interacció satisfactòria —per exemple, trobar un interlocutor també competent i poder fer ús de la llengua. Però és clar que aquest individu també aprén de les interaccions amb eventuals interlocutors, és a dir, aprén la probabilitat de trobar individus competents o usuaris. De manera que, en una sèrie d'interacció aleatòries, el temps de trobar aquests individus tindrà a veure amb la relació no lineal de competències i usos que hem definit en l'índex de desigualtat. Per això, amb el nostre ratolí virtual el que fem és definir uns paràmetres que estan relacionats amb les variables sociolingüístiques d'una regió i controlar el temps d'aquella trobada exitosa. Però, naturalment, el seu comportament té un punt d'aleatorietat. Per la qual cosa, no cal definir una trobada, sinó un bon nombre de trobades —de la mateixa manera que si fem un únic llançament d'un dau, els valors de les seues cares són equiprobables en el resultat, però si en fem molts la mitjana dels resultats dels llançaments tendeix a la mitjana del valor segon les cares. En el cas de la xarxa neuronal, però, com que es tracta d'un procés d'aprenentatge —a diferència del dau, que no aprén—, es produeix un efecte de

saturació que podríem representar amb una corba logarítmica: cada vegada al ratolí virtual li costa menys arribar al seu objectiu. Per això tampoc té sentit que el nombre d'intents (*trials*) siga molt elevat. En tots els casos se n'ha fent 100 intents i el programa ha fet un recompte automàtic del temps que consumeixen. Lògicament, el resultat en temps no és un valor absolut, sinó relatiu al nombre de neurones virtuals, les seues connexions o la grandària de l'espai virtual[5]. Per això, aquests variables s'elaboren de manera idèntica per a totes les regions. No obstant això, la qüestió important és la *correlació* entre els valors de l'experiment —que podem fer anàlegs a variables sociolingüístiques— i els de la desigualtat. Els resultats d'aquests ítems s'arrepleguen en la taula 4:

Taula 4

Regió	**Gamma**	**Èpsilon**	**Taxa apr,**	**Temps (100 intents)**
Reg. prov. Alacant	0,069	0,851	0,4658	6369
Reg. prov. Castelló	0,398	0,602	0,6376	2507
Reg. Alcoi-Gandia	0,547	0,453	0,7289	2045
Reg. València i àrea metr.	0,212	0,788	0,5432	3642
Reg. prov. València	0,543	0,457	0,7212	1733

Font: Elaboració pròpia.

Com seria previsible i encara que disposem de pocs valors, el temps resultant correlaciona de manera significativa i inversa amb gamma R = -0,946, de manera directa amb èpsilon R = 0,898 i de manera inversa amb la taxa d'aprenentatge R = -0,940. Ara bé, la conclusió més interessant és, sense dubte, la correlació entre el temps que consumeix el ratolí en finalitzar els seus 100 intents amb els paràmetres que hem indicat i l'índex de desigualtat

5. El mateix passaria si l'experiment el realitzara amb ratolins reals, per exemple, per recórrer laberints, amb un control estricte dels neurotransmissors: el temps total depén de la llargària del laberint, etc., i per tant els resultats s'expressen com a correlacions (o proves khi quadrat) entre quantitat de neurotransmissors i temps. Un avantatge indubtable de l'experimentació amb XNA és la possibilitat de poder controlar les tres variables indicades *al mateix temps*, mentre que en el cas de les xarxes biològiques només es pot controlar una.

lingüística que hem presentat anteriorment amb R = 0,747, amb un coeficient de determinitat de R^2 = 0,588.

4. Conclusions

Amb aquest article hem volgut presentar dues aplicacions de la IA per a la recerca sociolingüística del valencià perquè creiem que són escaients per a la revitalització de la llengua. Es podria pensar que la sociolingüística valenciana està ancorada en els mateixos conceptes forjats fa més de cinquanta anys, li manca modelització i moltes de les contribucions no van més enllà d'una glossa de percentatges. En aquest sentit, la IA suposa una oportunitat per fer un *aggiornamento* de la disciplina i oblidar definitivament nocions sense referents empírics clars. Som conscients que les aportacions que fem aviat quedaran obsoletes, pels avanços de la IA, no obstant això, es mantindrà la voluntat de perfeccionar la recerca sobre la nostra llengua. Amb tot i això, hem mostrat la virtualitat per a fer prediccions, la possibilitat de determinar un índex de desigualtat lingüística o de simular comportaments lingüístics amb XNA, amb el resultat summament satisfactori d'un coeficient de determinitat de R^2 = 0,588.

Hem pogut comprovar que les dues metodologies emprades —mineria de dades i xarxes neuronals— es poden relacionar i hem arribat a una conclusió summament interessant: amb els paràmetres que hem definit existeix una elevada correlació entre la desigualtat i el temps del model virtual. Altrament dit, hem emprat la simulació neuronal per verificar un índex en el qual hem fet servir la mineria de dades. Podríem anar encara més enllà. En aquest article hem emprat un model simple de xarxes neuronals. Podríem fer servir models més complexos. I encara més: no només podríem aplicar els resultats de la mineria de dades a les xarxes neuronals, com hem fet ací, sinó que també podríem aplicar els resultats de les XNA a la mineria de dades. Per exemple, ensinistrant l'algorisme de Random Forest per tal que ens permetera fer més prediccions sobre els resultats que hem vist —taula 4—, prediccions per a experiments no realitzats, i en aquest sentit avançar en la modelització comentada al principi.

Bibliografia

Álvarez Enparantza, Jose Luis (Txillardegi). (1984). *Elebidun gizarteen azterketa matematikoa*. Iruñea: UEU. Disponible en línia a: <https://www.jakin.eus/show/cfd5c0dbbcd0c927fb879b55f136ca87d815d-4bb>

Balaguer, Pau et al. (2019). CatSent: a Catalan sentiment analysis website. *Multimedia Tools and Applications*. 78: 28137-28155. https://doi.org/10.1007/s11042-019-07877-7

Breiman, Leo. (2001). "Random Forests". *Machine Learning* 45, 5-32. https://doi.org/10.1023/A:1010933404324

ECUV (2005). Enquesta 2005. Sobre coneixement i ús social del valencià (síntesi de resultats) [en línia]. València: Generalitat Valenciana. Conselleria d'Educació, Investigació, Cultura i Esport.

ECUV (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.

ECUV (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.

ECUV (2021). Enquesta 2021. Sobre coneixement i ús social del valencià (síntesi de resultats) [en línia]. València: Generalitat Valenciana. Conselleria d'Educació, Investigació, Cultura i Esport.

Fabà, Albert (2003). L'ús interpersonal del català i altres variables sociolingüístiques. Assaig d'un model interpretatiu. El cas de Santa Coloma de Gramenet. *Revista de Llengua i Dret*, núm. 40, p. 185-229.

Gupta, Sarah & DiPadova, Anthony. (2019). Deep Learning and Sociophonetics: Automatic Coding of Rhoticity Using Neural Networks. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 92-96, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-3013

Freund, Yoav i Schapire, Robert. (1996). Experiments with a New Boosting Algorithm. International Conference on Machine Learning, Bari, 3-6 July 1996, 148-156.

Hernàndez F. J. (2016). *El tio Canya ha mort. Notes sobre la mecànica sociolingüística del valencià*. València: Fundació Nexe. Disponible en: https://fundacionexe.org/publicacions/el-tio-canya-ha-mort

Hernàndez F. J. (2020a). La Relació entre competència (oral activa) i ús (públic): Un model matemàtic. *Treballs de sociolingüística Catalana*, núm. 30, juliol, 235-48.

Hernàndez F. J. (2020b). Adaptació del coeficient de GINI per a la formulació d'un coeficient de desigualtat lingüística. *Arxius de sociologia*, núm. 42, 235-263. https://doi.org/10.36950/elies.2020.42.8480

Hernàndez F. J. (2023). Formulació d'un indicador unitari de competència i d'ús de la llengua per a l'avaluació de polítiques lingüístiques. *Treballs de sociolingüística Catalana*, núm. 33, 141-157.

Hernàndez, Marc i Hernàndez Francesc Jesús. (2022). *Translatum*. Informe de recerca presentat a la Conselleria d'Educació, Universitats i Ocupació (inèdit).

Kretzschmar, William A. Jr (2008). "Neural networks and the linguistics of speech", *Interdisciplinary Science Reviews*, núm. 33: 4, 336-356. https://doi.org/10.1179/174327908X392898

Piloto, Luis S. et al. (2022). "Intuitive physics learning in a deep-learning model inspired by developmental psychology". *Nature Human Behaviour*, 11 de juliol <https://www.nature.com/articles/s41562-022-01394-8>.

Rodríguez, Pablo. (2018). *Inteligencia artificial*. Barcelona: ed. Deusto.

Simões, Alberto et al. (2014). Language identification: a neural network approach < https://hdl.handle.net/1822/30676>.

Sutton, Richard S. i Barto, Andrew G. (2014). *Reinforcement Learning: An Introduction* [Second edition in progress, 2014, 2015]. Cambridge, Massachusetts; Londres: MIT Press.

Thomas, Erik R.: "Sociolinguistic variables and cognition". *WIREs Cognitive Science* 2: 16: 701-716. https://doi.org/10.1002/wcs.152

Torrijos, Carmen i Sánchez, José Carlos. (2023). *La primavera de la Inteligencia Artificial*. Madrid: La Catarata.

Tosi, Zach i Yoshimi, Jeff. (2016). "Simbrain 3. A Flexible, Visually-Oriented Neural Network Simulator". *Neural Networks* (4 de juliol). Disponible en: https://www.sciencedirect.com/science/article/pii/S0893608016300879 https://doi.org/10.1016/j.neunet.2016.07.005

Tsoukala, Chara et al. (2020). "Simulating Code-switching Using a Neural Network Model of Bilingual Sentence Production". *Computational Brain & Behavior*. 4: 87-100. https://doi.org/10.1007/s42113-020-00088-6

Yoshimi, Jeff; Tosi, Zoë; Hotton, Scott; Gordon, Chelsea i Noelle, David C. (2023). *Neural Networks in Cognitive Science*. Versió 2023.1.1 . Disponible en: <http://www.simbrain.net/Documentation/v3/SimbrainDocs.html>