

# Euskarazko kulturgintzaren presentzia Euskal Herriko hedabide digitaletan: euskal kulturgileen identifikazioa NLP eta Adimen Artifizialeko eredu neuronalen bitartez

behategia@behategia.eus  
www.behategia.eus  
946018154

Libe Mimenza Castillo, Ane Martinez Juez, Naroa Burreso Pardo (NOR ikerketa taldea, UPV/EHU)  
Xabier Saralegi Urizar, Asier Garin Aldezabal (Orai NLP Teknologia)

Euskarazko kulturgintzaren gaur egungo presentzia mediatikoa zer nolakoa den aztertu dugu azken ikerketan, euskarazko kulturgileak zenbat agertzen diren eta nortzuk agertzen diren behatuz.

Emaitzen analisiaren ondorio nagusiak hurrengoak dira:

- Euskarazko hedabideek eta elebidunek kultur paisaia ezberdinak irudikatzen dituzte; euskaraz argitaratzen diren edukiak dira euskarazko kulturgileen ikusgarritasuna gehien lantzen dutenak.
- Harreman adierazgarria frogatu da: euskara hutsezko hedabideek euskarazko kulturgile gehiago agertzen dituzte, diziplina guztietan gertatzen da hori.
- Gakoa hizkuntzan dago: euskarazko edukietan euskarazko kulturgileen presentzia handiagoa agertzen da; baita hedabide elebidunetan ere, euskara hutsezkoen pareko agertuta zenbait kasutan. Gaztelaniazko edukietan, aldiz, euskaraz sortzen duten kulturgileen agerpen-maila baxuagoa da.

Beste emaitza aipagarri bat **Euskal Herriko 5.637 kulturgileen datu-basea sortu dugula da**. Horretarako, hurrengo prozesua jarraitu dugu:

## LAGINAREN PRESTAKETA

Q after:2022-12-31 before:2024-01-01 in: /kultura

Hamaika agerkitako «kultura» ataleko edukien izenburu eta gorputzak biltzeko eta formatu garbian ezarteko «boilerplate removal» izeneko teknikak erabili dira. Prozesua bi pausotan egin da garbiketa-kalitate gorena bermatzeko: lehenengo *Trafilatura* tresna erabili da, eta ondoren ad-hoc-eko garbitzaileak implementatu dira lehenak ondo garbitzen ez zituen iturrietarako.

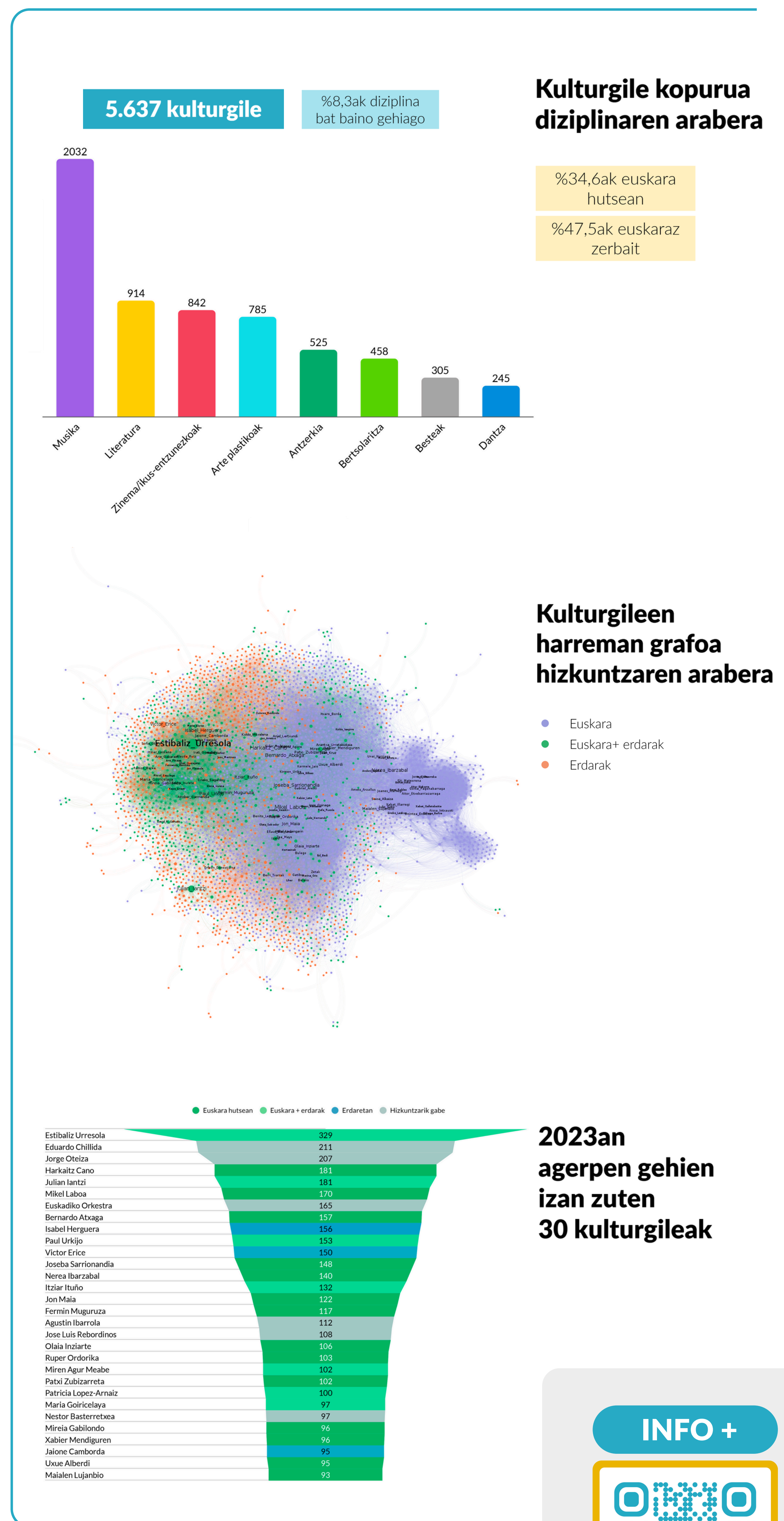
## EUSKAL KULTURGILEEN ERAUZKETA ETA PROZESAKETA

- Eduki bakoitzetik euskal kulturgile hautagaiak identifikatu dira **Meta-Llama-3-70B-Instruct hizkuntza-eredu handia (LLM)** erabiliz.
- Datu-basearen **garbiketa, bateraketa eta normalizazioa** egin da «string similarity»-aren arabera multzokatzean eta erregela ortotipografikoetan oinarrituta.
- DIZIPLINA** eta **JARDUTEN DUTEN HIZKUNTZA** automatikoki esleitzeko, ezagutza-base bat sortu da kanpoko ezagutza-iturriak erabilita: hiztegi horren bidez identifikatutako 13.239 entitateetatik **%13,4ari automatikoki esleitu zaizkio** bi aldagai horien atributoak (eta ondoren eskuz balioztatu dira).
- Kodetzea eskuz** egin da, eta kodetze-gida bat osatuz eta hainbat kontraste saio eginez bermatu da sailkapenaren koherentzia.
- Sailkatu direnen artean, % 40,9 kanpoko kulturgile identifikatu eta ezabatu dira eta gainontzekoekin osatu da Euskal Herriko 5.637 kulturgileko datu-basea.

## DATUEN ANALISIA

Euskarazko kulturgintzaren presentzia mediatikoa zer nolakoa den aztertzeko datu-basea osatu eta balioztatutakoan, honako datuen analisia egin da:

- Analisi estatistiko deskribatzailea eta inferentziala**, aldagaien arteko harremana existitzen ote den ezagutzeko Khi karratuaren hipotesi kontrastea erabilita.
- Euskal kulturgileen arteko **erlazioen analisia Force Atlas 2 bisualizazio-algoritmoa erabilita**.



Laguntzaileak:

