Visualització de dades

PRA1 – Selecció del conjunt de dades

Noms:Carlos Martínez Torró ([cmtorro@uoc.edu](mailto:cmtorro@uoc.edu))

Conjunt de dades escollit**:** [La llengua catalana als municipis de Catalunya](https://analisi.transparenciacatalunya.cat/es/Societat-benestar/-catal-947-La-llengua-catalana-als-municipis-de-Ca/ct77-e63k)

1. Justifiqueu breument la vostra selecció, sigui per motius personals o professionals.

Des de fa uns anys, hem entrat en una era on la creació de contingut multimèdia ha passat de fer-se majoritàriament als mitjans de comunicació tradicionals a realitzar-se per Internet. En aquest context, l’anglès i el castellà s’han fet amb un posició predominant i gran part del contingut que consumim a Catalunya i a la resta de l’Estat és en aquests idiomes.

Personalment, opino que aquesta posició de força que té en l’actualitat el castellà està afectant a l’ús del català. Tinc la sensació que el català es troba en retrocés en pràcticament tots els àmbits. Especialment, em preocupa que els parlants més joves (els qui han de sostenir la llengua en el futur i alhora el públic preferent de molts creadors de contingut a Internet) utilitzin cada cop menys el català en el seu dia a dia. Crec que és un tema particularment interessant i amb conseqüències de gran transcendència.

Per aquests motius, m’agradaria endinsar-me en aquestes dades per a veure com és la situació del català als diferents municipis i comarques del territori i, d’aquesta forma, polsar l’estat actual i passat de la llengua.

1. La rellevància del conjunt de dades en llur context. Són dades actuals? Tracten un tema important per algun col·lectiu concret? S’ha tingut en compte la perspectiva de gènere?

[Les dades formen part del projecte #catala947 del Departament de Cultura de la Generalitat](https://llengua.gencat.cat/ca/serveis/dades_i_estudis/poblacio/catala947/) i es van actualitzar per darrer cop al febrer de 2023. Les últimes dades al conjunt són de l’any 2019, tot i que [hi ha el compromís d’actualitzar-les en el futur](https://llengua.gencat.cat/web/.content/documents/dadesestudis/informes-locals/catala947.pdf). És un tema força rellevant per a tots els habitants de Catalunya, ja que la llengua impacta en el seu dia a dia i té un valor cultural incalculable. A més, aquestes dades també poden ser útils en altres àmbits, com el polític i l’acadèmic, per a diagnosticar l’estat de la llengua i poder dissenyar mesures per promoure i protegir la llengua.

Pel que fa a la perspectiva de gènere, només en un dels subconjunts de dades es té en compte el sexe dels habitants, així que no hi ha informació particular sobre l’ús de la llengua en funció del gènere. Aquesta podria ser una millora potencial en futures actualitzacions de l’estudi, ja que aquesta informació segregada per gènere podria aportar noves perspectives.

1. La complexitat (mida, variables disponibles, tipus de dades, etc.). Té de l’ordre de centenars o milers de registres? Té de l’ordre de desenes de variables? Combina dades categòriques i quantitatives? Inclou altres tipus de dades? Eviteu els conjunts excessivament simples.

El conjunt de dades es troba en un fitxer Excel (.xlsx) que conté 21 fulls: el primer és un índex i els altres 20 són diferents subconjunts de dades organitzats per seccions (**Taula 1**). El nombre d’observacions i variables canvia segons les dades concretes de cada full, però en molts d’ells les observacions són centenars i en alguns casos milers. Les variables també canvien segons el ull, però en general trobem una bona quantitat de variables. La majoria d’elles són quantitatives, ja que en molts casos tenim xifres d’habitants o de percentatges, mentre que entre les variables categòriques el més freqüent és trobar el nom del municipi, comarca o àmbit territorial, tot i que també en podem trobar d’altres com el sexe o el lloc de naixement. A més, enriquiré els diferents subconjunts de dades [amb les dades geoespacials de cada municipi](https://territori.gencat.cat/ca/06_territori_i_urbanisme/observatori_territori/mapa_urbanistic_de_catalunya/serveis_web_i_dades_obertes/descarrega-de-dades/format-shapefile-shp/), proveïdes pel Departament de Territori, amb l’objectiu de poder representar les dades en funció del seu context territorial.

Taula 1. Informació disponible al fitxer Excel del projecte #catala947.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dades** | **Full d’Excel** | **Observacions** | **Variables** |
| 2 Població segons el lloc de naixement 2000-2019 | dades\_2\_1 | 18940 | 5 |
| 2 Població segons sexe, edat i lloc de naixement 2019 | dades\_2\_2\_1 | 5682 | 22 |
| 2 Població segons lloc de naixement 2019 | dades\_2\_2\_2 | 947 | 16 |
| 3 Coneixement de català. 1991-2011 | dades\_3\_1 | 3780 | 8 |
| 3 Coneixements de català àmbit territorial i Catalunya | dades\_3\_2 | 21 | 6 |
| 3 Llengües que sap parlar fluidament | dades\_3\_3 | 21 | 5 |
| 3 Coneixements de català segon el lloc de naixement | dades\_3\_4 | 19 | 14 |
| 4 Llengua inicial, d'identificació i habitual | dades\_4\_1 | 21 | 23 |
| 4 % d'ús de català en un dia qualsevol | dades\_4\_2 | 19 | 9 |
| 5 Evolució de les inscripcions al CPNL | dades\_5\_1 | 219 | 9 |
| 5 Inscripcions segon el nivell | dades\_5\_2 | 219 | 4 |
| 5 Inscripcions segon origen | dades\_5\_3 | 219 | 4 |
| 5 Inscripcions segons origen estrangers | dades\_5\_3\_1 | 219 | 8 |
| 5 Parla.cat | dades\_5\_4 | 595 | 3 |
| 5 Voluntariat per la Llengua | dades\_5\_5 | 621 | 6 |
| 6 Índex Ofercat retolació segon sectors | dades\_6\_1 | 19 | 22 |
| 6 Índex Ofercat llengua oral segon sectors | dades\_6\_2 | 19 | 22 |
| 6 Índexs Ofercat retolació ciutats | dades\_6\_3 | 27 | 5 |
| 6 Índexs Ofercat llengua oral ciutats | dades\_6\_4 | 27 | 5 |
| 6 Adequació a la LPL | dades\_6\_5 | 31 | 12 |

En resum, parlem d’un conjunt de dades ric i complex on es poden generar diferents subconjunts de dades ajuntant les dades dels fulls que comparteixen secció. Aquest serà un pas crític, ja que haurem de garantir la coherència de les dades que estem unint en cada pas.

1. L'originalitat. No repetiu els conjunts de dades clàssics. Podeu, però, combinar-ne o millorar visualitzacions existents. Així, hi ha altres visualitzacions basades en aquest conjunt de dades? És una evolució o actualització d’un conjunt anterior? Heu enriquit un conjunt de dades ja existent?

Personalment, crec que és un anàlisi original, ja que no he pogut trobar cap visualització que empri aquestes dades. A més, la complexitat del conjunt de dades permet aproximacions diferents a les dades: no parlem del típic fitxer únic en format CSV amb X observacions i Y variables, sinó que hi ha 20 fulls amb dades que es poden utilitzar de forma independent o combinar en alguns casos. Així, és molt difícil que la meva aproximació sigui la mateixa que la que podria tenir un altre company; i, per altra banda, també podríem prioritzar de forma diferent els continguts del projecte. Per últim, l’enriquiment amb les dades geoespacials de cada municipi ens pot aportar una perspectiva més àmplia de l’ús del català segons l’àmbit geogràfic.

Per tant, la riquesa i actualitat del conjunt de dades unides a una rellevància més aviat local d’aquestes dades són dos factors que garanteixen l’originalitat de l’estudi que portarem a terme a la Pràctica 2.

1. Les qüestions que respondreu amb la visualització de dades, tenen en compte els punts anteriors? Estan ben plantejades? Són adequades pel conjunt de dades triat?

Considero que he triat les dades atenent-me a tots aquests punts i fent un especial èmfasi en l’originalitat de l’estudi. Inicialment, quan vaig veure fa uns mesos que hauríem de fer aquesta pràctica, vaig pensar en fer-la sobre el clima (un tema de moltíssima actualitat) o l’epidèmia de COVID (de menys actualitat, però de gran rellevància per la Ciència de Dades). No obstant, vaig acabar descartant tots dos temes per la seva manca d’originalitat i perquè era probable que ja hi hagués companys que s’haguessin decidit per ells. Tot i que són temes prou amplis i que es poden abordar des de molts punts de vista, el risc de coincidència era notable.

Així, crec que aquest conjunt de dades és prou original i ric com per poder plantejar múltiples preguntes d’acord amb els punts anteriors. A més, l’addició de les dades geogràfiques podria donar-nos informació sobre algun patró a les dades que, sense la informació geoespacial, seria molt difícil de detectar. Per tant, tot i que després amb l’exploració més en profunditat de les dades podran sorgir altres preguntes, ja ens podríem fer algunes qüestions amb aquestes dades:

* Quins són els municipis i les comarques on més es parla català? I, sobre aquesta pregunta, ens podríem fer alguna més:
  + Hi ha relació amb l’origen dels seus habitants?
  + És també així en el seu context geogràfic més proper?
  + Quina ha estat la seva evolució en els darrers anys? (Tenim informació de 1991, 1996, 2001 i 2011 a banda de 2019).
* Quin és el domini del català de la gent no nascuda a Catalunya en comparació amb els nadius? Canvia segons el seu àmbit territorial?
* Relacionat amb això, hi ha diferències en l’ús del català com a llengua habitual en cada àmbit territorial? Depèn de l’origen de la persona?
* Quina és l’evolució de les iniciatives per promoure l’ús del català?

Aquestes són algunes de les preguntes que considero rellevants per tenir una idea de l’estat actual del català i que es podrien respondre amb les dades que hi trobem en aquest conjunt.