

I Jornades de Llengua i Estadística

20 i 21 de novembre del 2008

Ponències presentades

Barcelona, desembre del 2009

Idescat

Edita:

Generalitat de Catalunya

Institut d'Estadística de Catalunya

Via Laietana, 58

08003 Barcelona

<http://www.idescat.cat>

1a. edició: Barcelona, desembre del 2009

ISBN: 978-84-393-8093-1

ISSN: 2013-3847 versió impresa

2013-5157 versió digital

Dipòsit legal: B-45.597-2009 versió impresa

B-45.598-2009 versió digital



Aquesta obra està subjecta a una llicència de Reconeixement-No comercial-Sense obres derivades 3.0 Espanya de Creative Commons.

<http://creativecommons.org/licenses/by-nc-nd/3.0/es/deed.ca>

La versió digital inclou els canvis que s'hagin pogut produir després del tancament de l'edició en paper.

<http://idescat.cat/p/quad02>

Consell editorial: Joan Abril, Txema Castiella, Gemma García, Marta Masats, Josep Missé, Enric Ripoll i Cristina Rovira

Presentació

L'Institut d'Estadística de Catalunya va organitzar el 2008 les Primeres Jornades de Llengua i Estadística, que van reunir una comunitat formada per economistes, estadístics, lingüistes, matemàtics, periodistes, sociòlegs... Les ponències que publiquem representen les diverses àrees científiques en què es van estructurar aquestes Jornades i representen també la plena dedicació a pensar i repensar la difusió estadística més enllà de la presentació extremadament calculada d'uns resultats estadístics, d'un qüestionari o del procés d'una operació estadística.

Les Jornades van tenir com a objectiu conèixer la interdependència de les dues disciplines estadístiques que encapçalen el títol: la llengua i l'estadística o l'estadística i la llengua, amb un ordre de factors que no altera el producte, que és comunicar dades però amb la capacitat suficient per difondre-les correctament, és a dir, amb accessibilitat i disponibilitat, dos valors que pertanyen al Codi de bones pràctiques estadístiques i que són, en bona part, el fil conductor d'aquestes ponències.

Les activitats estadístiques, els mitjans de comunicació i la societat són l'escenari en què es mouen aquests dos grans protagonistes. Un escenari, d'altra banda, on la terminologia, el tractament informatiu de les dades o l'accessibilitat social tenen punts de confluència i de reflexió en la difusió estadística. A un any de la celebració, els debats i els resultats d'aquestes Jornades han ajudat i ajudaran a millorar el nostre servei a la societat, oferint informació útil i rigorosa, alhora que precisa i accessible.

Anna Ventura i Estalella

Directora de l'Idescat

Comitè científic

Institut d'Estudis Catalans

Joan Martí

(president de la Secció Filològica i antic rector de la Universitat Rovira i Virgili)

Departament de la Vicepresidència

Montserrat Romà

(cap del Servei d'Informació, Difusió i Estudis de la Secretaria de Política Lingüística)

Centre de Terminologia TERMCAT

Rosa Colomer (directora)

Glòria Fontova (terminòloga i cap de l'Àrea d'Assessorament)

Universitat Oberta de Catalunya (UOC)

Miquel Strubell (professor d'Estudis de Llengües i Cultures)

Idescat

Joan Abril (assessor lingüístic)

Eva Cabrera (assessora lingüística i secretària del Comitè)

Txema Castiella (subdirector de Difusió Estadística)

Josep M. Martínez (estadístic responsable d'operacions)

Marta Masats (cap de l'Àrea d'Estadístiques Socials)

Anna Ventura (directora)

Índex

Presentació	3
Institucions col·laboradores i programa	4
Introducció. Salvador Giner.....	7
Antoni Castells.....	9
Ponències	
Ponència inaugural	
La terminologia i el rigor en l'estadística	
Eduard Bonet.....	11
Àrea de Llengua i Activitat Estadística	
1 Els llenguatges estadístics i la llengua catalana	
Dolors Presagué Vila.....	25
2 Recursos terminològics per al treball estadístic	
Glòria Fontova.....	37
3 El lenguaje estadístico y el inglés	
Francisco Hernández Jiménez	43
4 El miedo al botijo produce monstruos	
Joxean Sagastizabal	51
5 Cuestiones de difusión estadística relativas al gallego	
Carlos Iglesias/Esther Vizcaíno.....	55
Àrea de Llengua i Mitjans de Comunicació	
1 La lectura de les dades estadístiques en els mitjans de comunicació	
Albert Closas	63
2 Vics i defectes del llenguatge estadístic en els mitjans	
Josep Maria Ureta	67
3 L'adaptació de les notes de premsa estadístiques en els mitjans	
Joan Roy	69
4 Com presentem la informació estadística a Internet	
Xavier Rius	73

Àrea de Llengua i Societat

1	Disponibilitat lingüística en països plurilingües	
	Guillem Rico	75
2	Esforç institucional per un llenguatge accessible	
	Joan Abril.....	81
3	La llengua com a indicador i instrument d'una operació estadística	
	Elena Sintès	89
4	L'estadística i la llengua com a elements de cohesió social: l'anàlisi de l'estructura social catalana amb les dades del PaD	
	Cristina Sánchez	97

Ponència de tancament

	Mesurem els mots? L'anàlisi estadística de textos	
	Mònica Bécue	119

Cloenda

	Bernat Joan	137
--	-------------------	-----

Introducció

Salvador Giner

President de l'Institut d'Estudis Catalans

Bon dia a tothom, honorable conseller, senyores i senyors. Benvolguts companys, amics de l'Institut i participants a les Jornades de Llengua i Estadística. Em plau moltíssim rebre el senyor conseller, amb un tema importantíssim, en aquesta acadèmia nacional de les ciències i de les humanitats, que és casa vostra.

És un plaer especialment, al marge del meu càrrec en aquesta casa, veure l'aplicació de l'estadística a la llengua, d'una disciplina, inventada per un metge belga el 1850 aproximadament, que va esdevenir sociòleg i la va batejar amb el nom d'estadística. Hi havia el nou estat belga, constituït el 1835 (una mica artificial segon diuen els mateixos belgues). I el cas és que aquest descobriment, l'estadística, la ciència de l'estat, dades sobre l'estat, sobre suïcidis, collites, distribució de la renda, d'una manera prou primitiva, ha estat un dels descobriments o dels invents, si voleu, de les disciplines de les ciències socials, que més influència ha tingut sobre la ciència en el món modern.

S'ha escampat primer, naturalment, a la demografia i a l'economia, d'aquella primitiva sociologia, i molt poc després a la medicina, a la biologia, a totes les disciplines, a la física quàntica i a tot el que us pugueu imaginar.

La lingüística, naturalment, i la llengua i la filologia no són pas alienes a això, i és per aquesta raó que celebrem que es realitzi una jornada com la d'avui, per relacionar la llengua amb l'estadística. Estem mesurant la nostra llengua catalana i totes les altres llengües, però la nostra especialment, de la qual tenim cura.

L'única manera de tractar-la i conèixer-la bé és fent servir les eines de l'estadística, i també, ara ho discutia amb el conseller, llançar hipòtesis causals i explicatives dels processos lingüístics. Això és tot el que vull dir, res més, i donaré la paraula a l'honorable conseller Antoni Castells.

Moltes gràcies.

Antoni Castells

Conseller d'Economia i Finances de la Generalitat de Catalunya

Moltes gràcies, senyor president de l'Institut d'Estudis Catalans, senyora directora general de l'Idescat, amics tots.

Vull agrair a l'amic Salvador Giner l'amabilitat que ha tingut d'acollir-nos en aquesta casa, que és la casa de tots i que simbolitza l'aspiració a l'excel·lència d'aquest país, tant en el terreny cultural com en el terreny científic.

Ja fa uns 100 anys, potser una mica més, que va ser creat aquest Institut d'Estudis Catalans. I avui l'Institut acull una iniciativa com és aquestes jornades sobre llengua i estadística. El primer que haig de fer és felicitar els que han tingut l'encert d'impulsar-la: l'Idescat, en primer lloc, l'Institut d'Estudis Catalans, el TERMCAT, la UOC, la Secretaria de Política Lingüística... Tots hi han col·laborat.

Poder ser avui aquí amb tots vostès, en la inauguració d'aquestes jornades sobre llengua i estadística, és una d'aquelles gratificacions que proporciona ser conseller d'Economia i Finances i, per tant, responsable últim de l'Idescat. I estirant el fil de l'Idescat arribem a la llengua, i de la llengua a l'Institut d'Estudis Catalans.

Llengua i estadística. Tots dos són instruments que serveixen per comunicar-nos. La llengua, no cal dir-ho, i també l'estadística, que no deixa de ser un instrument, un llenguatge, si voleu. El llenguatge de convertir en indicadors, en números, una realitat molt àmplia. L'estadística ens serveix per comunicar-nos, per saber de què estem parlant en moltíssims àmbits. Les estadístiques són un instrument, però també són una expressió de la realitat. Són les dues coses. La llengua també és un instrument. Sí que ho és, és clar que ho és. Si no és un instrument per comunicar-nos, no serveix. Però alguns la voldrien reduir només a un instrument i és més que això: és una expressió de nosaltres mateixos.

I l'estadística igual. L'estadística és un instrument, és clar que sí. Però si no és capaç d'expressar la realitat no compleix la seva funció, per molt que sigui un instrument. De manera que les dues, llengua i estadística, són instruments per comunicar-nos. I això vol dir que quan les fem servir, el que emet el missatge i el que el rep han d'entendre's.

Però, al mateix temps, també són una expressió de la realitat, o sigui que si no són capaces de reflectir la realitat, malament. I no sempre és fàcil, això.

El llenguatge de l'estadística és complicat, de vegades. Sovint diem: "El PIB ha crescut un X%", o bé: "La productivitat de la nostra economia és aquesta", o diem: "l'IPC és de l'X%", i quan ho diem aquí, a Catalunya, no estem dient el mateix que a altres llocs, perquè els conceptes utilitzats o la metodologia de càlcul, no són ben bé similars. De manera que és bo que les estadístiques serveixin per a les dues coses. Per comunicar-nos i per reflectir bé la realitat, de la manera més ajustada possible.

En el nostre cas, la vinculació entre llengua i estadística adquireix unes connotacions especials. Quan avui celebrem aquestes jornades, ho fem amb el propòsit d'afirmar que es poden fer esta-

dístiques en català, i bones estadístiques en català. Algú dirà: “Veieu aquests catalans? L’únic del que es preocupen és del català”. I nosaltres els hem de respondre: “Ens preocupem del català i de fer bones estadístiques en català. De les dues coses”. El nostre propòsit no és fer estadístiques en català encara que siguin dolentes. Fer estadístiques en català, sí, però les millors estadístiques. Aquest és el nostre propòsit; i a més, hem de demostrar-ho més que altres que són bones.

La terminologia i el rigor en l'estadística

Eduard Bonet

Catedràtic emèrit de la Universitat Ramon Llull a ESADE

Membre emèrit de l'Institut d'Estudis Catalans

PRIMERA PART

NOMS I CONCEPTES

Fixar i emprar correctament la terminologia d'una disciplina acadèmica són activitats que contribueixen fortament a establir i mantenir el rigor científic. Naturalment, hi ha altres aspectes metodològics que són necessaris per aconseguir que els desenvolupaments teòrics i les aplicacions pràctiques siguin rigorosos; però aquesta exigència no ens permet deixar de banda les qüestions de llenguatge. En el cas de l'estadística, les subtilitats lògiques del raonament i l'extensió social i mediàtica de les aplicacions fan que els problemes relacionats amb la comprensió del seu vocabulari portin a malentesos greus, tant en la interpretació de les dades de temes d'actualitat com en la de moltes recerques científiques. Per encarar aquests temes, crec que no serà ociós que tots plegats recordem la diferència que hi ha entre els significats dels noms en la vida quotidiana i en la ciència, i la que hi ha entre introduir un nom i definir un concepte.

El significat dels noms en l'ús ordinari del llenguatge

Una paradoxa i un problema sobre el significat dels noms en la vida quotidiana és que molts d'ells normalment els fem servir d'una manera correcta, però si ens demanen que en precisem o en definim el significat no som capaços de donar-hi una resposta satisfactòria. Per exemple, tots sabem què vol dir "aprendre" però, si demanem a la gent que ens ho expliqui, tindrem molta sort si ens contesten "és saber una cosa que no sabíem abans", o "és saber fer una cosa que no sabíem fer abans", ja que podem esperar que ens diguin coses de l'estil "aprenem a l'escola". Atès que la situació és aquesta, ens hem de preguntar com és possible que el llenguatge i la comunicació funcionin? Ens entenen perquè ens hem avesat a emprar un nom en un context determinat i, d'aquesta manera, n'establim i en comprenem el significat, però també estem exposats a molts malentesos.

Quan el context canvia, el significat dels noms també canvia. Per exemple, el mot *aprendre* s'ha relacionat estrictament amb les activitats escolars i, segons aquesta mentalitat, fa uns quants anys un director general d'una companyia s'hauria sentit ofès si li haguéssim dit que apren. Però, actualment, amb les idees de formació continuada durant tota la vida i les teories de l'aprenentatge corporatiu i de la gestió del coneixement, la noció d'aprenentatge ha adquirit un gran valor i considerem que una cosa important que fan els directius és aprendre i crear coneixement pràctic.

Ludwig Wittgenstein va tractar amb molta profunditat aquests temes en el llibre *Philosophical Investigations* (introducció i traducció al català de Josep Maria Terricabras, *Investigacions filosòfiques*, Editorial Laia, 1983; Edicions 62, 1997). En aquesta obra, que va constituir una revolució filosòfica, presenta un mecanisme bàsic que produeix canvis de significats i mostra que el significat dels noms en el llenguatge ordinari normalment no es pot definir en termes de condicions necessàries i suficients, com fem en les teories científiques. El seu argument segueix les línies següents: suposem que, d'entrada, un nom denota exclusivament una classe ben definida d'objectes i que, més endavant, s'hi introdueix una altra classe d'objectes que tenen semblances importants amb els de la primera. Una opció que tenim consisteix que, ateses les semblances, donem el nom dels objectes de la primera classe també als de la segona. D'aquesta manera, el significat d'aquest nom s'ha ampliat a través d'una metàfora. (Recordem que Aristòtil va caracteritzar la metàfora de la manera següent: una metàfora és una translació del nom d'una cosa a una altra cosa, en virtut d'una semblança o analogia que hi ha entre ambdues coses.)

Com que aquesta mena d'ampliacions del significat d'un nom es fan moltes vegades amb criteris de semblança diversos, el significat del nom cobreix, en expressió de George Lakoff, una categoria radial d'objectes que no es pot definir en termes de condicions necessàries i suficients. Wittgenstein va indicar que aquests objectes “tenen un aire de família” i crec que amb aquesta expressió al·ludia a les nostres capacitats cognitives per apreciar intuïtivament l'extensió d'un significat. Així, per Wittgenstein, el significat d'un nom depèn del context o, com ell va dir més acuradament, “del joc lingüístic en el qual l'empren”.

Des d'un punt de vista lògic, aquestes indefinicions, ambigüitats i multiplicitats de significats són un defecte greu de l'ús ordinari del llenguatge. Però, des d'un punt de vista cognitiu, es tracta d'una flexibilitat que ens permet expressar-nos i comunicar-nos, i que facilita els canvis de significats per adaptar-nos a noves situacions.

Definició del significat dels noms i dels conceptes en la ciència

Una exigència de la ciència és que els seus termes tinguin un significat precís i que els seus conceptes siguin ben definits. Definir un concepte és una operació lògica que no s'ha de confondre amb la de crear o introduir una terminologia, que és una operació lingüística. Tanmateix, com que definir un concepte comporta també donar-li un nom, aquestes activitats van molt juntes i, en molts textos, no queda clar si parlem d'una cosa o de l'altra. Atès que personalment en queda un cert deix platònic i aristotèlic més que nominalista, em concentraré en la definició de conceptes.

En les reflexions sobre conceptes, hem de considerar dues classes d'aspectes: els ontològics, que es refereixen a la seva naturalesa, i els epistemològics i lògics, que es refereixen a les maneres de definir-los. De les diverses doctrines filosòfiques que s'han presentat sobre aquests temes i que cobreixen quasi tota la història de la filosofia, n'indicaré algunes que podem prendre com a referència.

Sòcrates considerava que el coneixement i els conceptes, d'alguna manera, són a dins nostre i que la nostra feina és fer-los sortir enfora. D'acord amb aquesta manera de pensar, va introduir el seu mètode amb el nom de “maièutica”, que originalment es referia al treball de les llevadores quan ajuden les mares a treure el nen en el part. El mètode consistia en una sèrie de preguntes amb les quals volia fer veure als interlocutors que el coneixement i els conceptes eren més complicats que allò que es pensaven i amb les quals els portava a buscar la veritat. Les converses socràtiques consideraven que els conceptes existeixen independentment de nosaltres i no en negociaven la definició.

Plató va desenvolupar aquestes idees i les va portar a plans filosòfics molt més alts. Segons Plató,

les idees pures, com per exemple el concepte geomètric de línia recta, són eternes i immutables i constitueixen les essències de les coses. Així, les rectes que podem dibuixar són només imitacions imperfectes o pàl·lides ombres del concepte abstracte. L'essència de les coses resideix en les idees pures i no en els objectes que veiem. Aprendre és recordar. Per tant, definir un concepte o demostrar un teorema és recordar. El mètode per recordar consisteix en una sèrie de preguntes, que caracteritzarà la dialèctica.

Per recordar el mètode dialèctic que Sòcrates i Plató fan servir per definir un concepte, en comentaré i analitzaré d'una manera molt simplificada un exemple de la primera part del diàleg *Gòrgies*. Sòcrates pregunta a Gòrgies què és la retòrica i quin és el seu objecte, i ell li dóna de seguida una definició (l'objecte de la retòrica són els discursos). Sòcrates la critica (tots els discursos?: com els dels metges que indiquen als malalts què han de fer per guarir-se? És la medicina una art retòrica?) i empeny Gòrgies a modificar la definició original i proposar-ne una altra (la medicina no és una art retòrica perquè requereix accions físiques; la retòrica fa la seva feina només amb paraules). Aquest procés es repeteix diverses vegades fins que els interlocutors arriben a la definició correcta.

Al segle xx, els neopositivistes del cercle de Viena van considerar que un concepte i la seva definició és merament una abreviació d'una expressió llarga, és a dir, és una simplificació taquigràfica d'un sistema de signes en un llenguatge formal. Un exemple senzill, però no gaire rellevant, seria dir que el concepte de força és merament una abreviació de la fórmula massa $\times$ acceleració. Imre Lakatos, un seguidor de Karl Popper, matemàtic per formació i filòsof de la ciència per vocació, va reaccionar fortament contra aquella idea. Lakatos és ben conegut per la teoria que va presentar a l'obra *La metodologia dels programes de recerca científica*, que és d'una rigidesa increïble. Però en aquest text ens interessa una altra publicació, menys coneguda i molt més interessant.

El llibre de Lakatos *Proofs and Refutations* (1976) es pot llegir com el procés històric que van seguir les matemàtiques per definir el concepte de poliedre, el qual sembla molt intuïtiu. El mètode que el llibre posa en evidència per mitjà de diàlegs és molt semblant al dialèctic que he comentat i consisteix a proposar una definició inicial i anar modificant-la d'acord amb dos referents bàsics. D'una banda, el coneixement que ens dóna la nostra experiència dels cossos geomètrics i, en especial, de les formes complicades que tenen els minerals cristal·litzats, que queden maclats entre si. I, d'altra banda, els teoremes que les diverses definicions produeixen. És a dir, és una manera de treballar que té en compte el coneixement empíric i els desenvolupaments teòrics.

Des del punt de vista de la lògica, les definicions són convencionals i cauen fora de l'exigència de ser demostrades, la qual no es pot deixar de banda en les proposicions i els teoremes. Les definicions, per ser acceptables, no han de ser contradictòries i s'han d'establir a partir de conceptes definits prèviament (deixo de banda el tema dels conceptes inicials o primaris). Des del punt de vista epistemològic i cognitiu, els conceptes són peces bàsiques del pensament, que fan possible el desenvolupament i la comprensió de les teories. En definir un concepte, volem captar i formular les propietats que el caracteritzen, i moltes vegades ho fem, com en els exemples que he presentat, a partir d'examinar nocions del sentit comú i rigoritzar-les. Aquesta feina pot ser molt llarga i meticulosa (la definició del concepte de probabilitat es va discutir al llarg de més de dos segles). Tanmateix, els llibres científics normalment ens en donen les definicions sense cap explicació prèvia sobre com s'han discutit.

La creació de la terminologia científica

En la línia de pensament que estic exposant, la definició precisa del significat dels termes científics depèn de la definició dels conceptes corresponents que he comentat amb una certa extensió. La definició dels conceptes, per ser completa, requereix dotar-los d'un nom. Aquest tema ens porta al de la creació de la terminologia científica, que naturalment té una dimensió lingüística important.

La relació entre un nom i el seu significat també és convencional, en el sentit que no deriva de cap lligam entre les dues entitats, llevat del cas dels noms onomatopèics, que imiten un soroll o un crit. En contra de les tradicions màgiques que busquen el vertader nom secret de les divinitats, al qual atribueixen un gran poder, en principi podem posar a cada cosa el nom que ens agradi. Tanmateix, aquesta llibertat queda limitada per moltes regles. En primer lloc, els noms nous s'han d'adaptar a l'estructura de la llengua i han d'obeir les seves regles fonètiques i sintàctiques de composició i derivació. També han de tenir en compte la terminologia que històricament s'ha produït. En un altre pla conceptual, a cap persona no podem posar-li el nom "ningú", tal com Ulisses, l'heroi de tots els recursos, va dir que es deia per enganyar el ciclop Polifem, que havia tancat la seva tripulació en una cova i se l'anava menjant.

La creació de la terminologia científica requereix la participació de lingüistes i terminòlegs. En el camp de la medicina clàssica, els noms dels conceptes s'estableixen sistemàticament a partir de noms grecs, d'acord amb unes regles d'adaptació. En moltes disciplines, es prefereixen noms del llenguatge habitual, els quals passen a tenir, a més a més dels seus significats imprecisos, un significat tècnic ben definit. D'aquest procediment, en distingiré les formes següents. En el primer cas, els investigadors parteixen d'un concepte intuïtiu per definir un concepte científic i mantenen el nom del primer per referir-se al segon. Per exemple, la psicologia i la filosofia passen de la noció comuna al concepte científic d'aprenentatge sense canviar-ne el nom. En el segon cas, definim un concepte en el marc d'una teoria sense considerar cap concepte intuïtiu, però li donem un nom del llenguatge ordinari, de manera que el seu significat tingui alguna relació amb la noció nova. Per exemple, crec que, en estadística, l'expressió *mostra representativa*, que comentaré més endavant, es va originar d'aquesta manera. En tercer lloc, podem donar a un concepte científic un nom del llenguatge ordinari en virtut d'una analogia o semblança que hi ha entre els objectes dels dos conceptes. Es tracta d'una forma de creació de terminologia científica per mitjà de metàfores. Recordem una altra vegada la definició aristotèlica de metàfora. Seguint aquest procediment, una tècnica estadística que es pot aplicar a molts problemes va rebre el nom de *jack-knife*, terme que inicialment es refereix a una mena de ganivets de butxaca que tenen diverses peces i serveixen per fer moltes coses. També la tècnica de mostrejar va rebre el nom de *bootstrap*, que es refereix a la corretja que tenen les botes altes per ajudar-nos a posar-nos-les.

Unes aportacions de Gabriel Ferrater a la terminologia de l'estadística en català

La col·laboració de Gabriel Ferrater al llibre d'Eduard Bonet *Espais de probabilitat finits*, (Editorial Lavínia, 1969; Editorial Teide, 1972) va constituir una bona contribució a la normalització del llenguatge científic en català. Al llibre *Gabriel Ferrater: un amic, un temps, un món* (Publicacions de la Residència d'Investigadors, CSIC-Generalitat de Catalunya, 2008), n'explico alguns aspectes referents a l'estil, la sintaxi i la terminologia, dels quals recordaré alguns comentaris sobre els noms.

Els llibres en castellà empen les expressions *cota*, *conjunto acotado* i *función acotada*, que normalment traduïem al català fent servir el nom *cota*, que és correcte. Gabriel Ferrater va proposar les expressions *fita*, *conjunt afitat* i *funció afitada*, que va sostenir amb els arguments següents: en una carta que Ferrater havia escrit a Pere Puig Adam, segurament als anys cinquanta, es queixava que Puig Adam hagués traduït al castellà les expressions angleses *bound* i *bounded function*; les franceses *borne* i *fonction bornée*, i les alemanyes que no conec i que motivaven aquella protesta per les expressions *cota* i *función acotada*. Les expressions originals tenen el sentit de delimitació i de senyal de tanca, mentre que *cota* és refereix a l'altura d'un terreny, tal com els militars fan servir constantment en les maniobres. La traducció castellana correcta hauria de ser *coto*. Puig Adam li va donar la raó, però li va indicar que no era ni fàcil ni oportú fer el canvi. Poden dir que es tracta de dues metàfores i que totes dues encaixen bé amb la idea matemàtica que hi ha al darrere. En tot cas, la proposta de Gabriel Ferrater sobre *fita* i *funció afitada* va ser acceptada per la comunitat matemàtica catalana.

En estadística, el castellà emprà l'expressió *sucesos aleatorios* i, a l'Amèrica del Sud, *eventos aleatorios*, que provenen de la forma anglesa *random events*. En català normalment ho traduïem per *successos aleatoris*. Gabriel Ferrater va proposar l'expressió *esdeveniments aleatoris*. Segons ell, aquesta terminologia del càlcul de les probabilitats prové de l'expressió francesa *événements aleatoires*, en la qual el nom *événements* està relacionat amb el verb *devenir*, que té una connotació de futur que és important en la teoria de les probabilitats. En català, l'expressió *esdeveniments*, relacionada amb el verb *esdevenir*, manté la connotació de futur que és important no perdre.

Gabriel Ferrater també va remarcar la distinció entre “la puixança d'un conjunt” i “la potència d'un nombre”, fent servir noms diferents, distinció que, segons tinc entès però no puc confirmar, també es fa en alemany. Però el terme *puixança* no va tenir èxit. Quant al terme *varianza* en castellà, el va traduir per *variança* en català, però més endavant l'Institut d'Estudis Catalans el va normalitzar en la forma *variància*. Crec que el fet que l'Institut d'Estudis Catalans sigui alhora una acadèmia científica i una acadèmia de la llengua catalana facilita la creació i la normalització de la terminologia científica.

Funcions de la terminologia científica

La terminologia científica té dues funcions bàsiques que es complementen. La primera es relaciona amb la lògica a través de la definició precisa de conceptes. Els conceptes definits rigorosament permeten fer deduccions vàlides i demostrar teoremes. La segona es relaciona amb la comunicació i permet que els científics s'entenguin d'una manera eficaç. Tanmateix, el llenguatge científic té altres funcions que són importants en el camp social i en el camp cognitiu.

Fins als segles *xvi* i *xvii*, els treballs de teologia, filosofia i ciències s'escriuïen i es debatien en llatí. Aquest llenguatge universal, amb el qual els acadèmics es comunicaven, també era una barrera que limitava l'accés a aquestes disciplines a una elit relativament molt petita. L'única excepció important van ser els treballs de Ramon Llull (1232-1314) sobre filosofia, teologia, educació, lògica, retòrica i tota mena de temes. Ramon Llull es movia per un zel apostòlic extraordinari de convertir els infidels. Per aquesta raó, adreçava els seus arguments a tothom, desenvolupava els seus temes en nivells conceptuals diferents i escrivia i traduïa els seus llibres emprant el català, el llatí, l'hebreu i l'àrab. Galileo Galilei (1564-1642) és reconegut com el primer gran científic que va escriure molts treballs en romanç, en el seu cas en italià, amb la intenció que la ciència arribés a tothom. Aquesta actitud va irritar molts acadèmics i possiblement va contribuir, d'una manera indirecta, al procés de la Inquisició que el va condemnar. Actualment, la terminologia científica, especialment la que es desenvolupa a partir d'arrels gregues i llatines, contribueix a crear una mena d'elitisme. També el coneixement del llenguatge científic bàsic estableix unes diferències socials i culturals importants.

La terminologia científica creada a partir de metàfores té unes funcions cognitives importants, per raó de les dues propietats següents: d'una banda, es tracta d'un joc de paraules basat en una semblança o analogia entre dues coses. D'altra banda, l'analogia és un instrument bàsic, tant en les obres poètiques com en la creació científica. Així, per exemple, els antics egipcis van descobrir la fórmula del volum d'un cos cònic a partir de la metàfora següent: un con és una piràmide que té infinites cares i, per tant, el seu volum també és l'àrea de la base per l'altura dividit per tres. Aquesta fórmula no es va demostrar rigorosament fins al segle *xvi*. En la matemàtica moderna, la definició segons la qual un espai vectorial és un conjunt on hi ha definida una operació suma de dos elements i una operació producte d'un element per un nombre real, de manera que aquestes operacions compleixen determinades propietats, és una metàfora basada en l'analogia que hi ha entre aquests espais i el conjunt de vectors geomètrics. Aquesta metàfora ens permet pensar alguns espais de funcions com espais multidimensionals de vectors clàssics o segments orientats. Així, ens trobem amb una associació d'idees molt productiva que va portar a descobrir molts teoremes.

SEGONA PART

MALENTESOS EN L'ÚS D'ALGUNS TERMES ESTADÍSTICS

Com en les altres ciències matemàtiques, la terminologia estadística s'ha creat per donar noms a conceptes ben definits, a partir dels quals s'han demostrat molts teoremes, i per facilitar la comunicació entre els científics, i entre ells i la societat. Tanmateix, l'ús d'aquesta terminologia porta moltes vegades a malentesos en el món social, en els mitjans de comunicació i en els àmbits de recerca. Les raons que expliquen aquest problema són, en part, els significats diferents que els noms tenen en la ciència i en el llenguatge comú, com ja he comentat en la presentació anterior, i, en part, la subtilesa de la lògica de la probabilitat, que presentaré a continuació.

Sobre la noció de mostra representativa

Atesa la importància que té la noció de *mostra representativa*, ens podem preguntar què pensa molta gent d'aquest concepte, del qual no en coneix cap definició estadística. L'adjectiu *representatiu*, associat al nom *representació* i al verb *representar*, té molts significats en el llenguatge ordinari, com els que es relacionen amb una imitació de la realitat en una representació teatral, amb una delegació política que inclou totes les tendències, amb un dibuix que presenta un objecte o amb un mapa que presenta, d'una manera reduïda, una regió geogràfica. Tots aquests sentits contribueixen a donar una idea imprecisa del concepte de mostra representativa i a dotar-lo alhora d'una impressió de familiaritat. Tanmateix, aquestes idees són insuficients i moltes vegades, com en el cas de l'exemple que comentaré a continuació, porten a greus malentesos.

La noció de *mapa* com una representació que manté una proporció o escala constant entre les distàncies reals i les del dibuix suggereix el sistema de mostres per quotes, que es defineix de la manera següent: de la població, en coneixem el percentatge de persones que tenen determinades característiques d'edat, sexe, professió, estat civil, dimensions de la població on viuen i moltes d'altres. La mostra, pretesament representativa s'escull amb el criteri que respecti aquests percentatges. En unes eleccions presidencials dels Estats Units, una gran empresa d'opinió pública va intentar predir els resultats a partir d'una mostra per quotes que incloïa molts milers i milers de persones. La predicció va ser un gran fracàs que pràcticament no donava cap possibilitat a Truman enfront de Dewey. Aquella experiència i moltes d'altres van fer repensar el concepte de mostra representativa.

Podem definir que una mostra és representativa d'una població respecte d'una propietat A, si el percentatge $p(A)$ d'elements de la població que tenen aquesta propietat i el percentatge $f(A)$ d'elements de la mostra que la tenen són iguals o difereixen en menys d'un marge o interval predefinit. D'aquesta definició, en podem remarcar dues conseqüències importants. En primer lloc, una mostra pot ser representativa d'unes propietats i no ser-ho d'unes altres. En les eleccions mencionades, la mostra per quotes era representativa de les propietats que van servir per dissenyar-la, però no dels resultats electorals que es volien predir. En segon lloc, la definició constitueix, per dir-ho així, una constatació "a posteriori" que no ens dóna cap criteri per escollir la mostra.

Si no coneixem el percentatge de la població, que és la dada que volem predir o aproximar per mitjà de la mostra, d'entrada sembla impossible que hi hagi un criteri per triar la mostra. Tanmateix, aquest problema té una solució a través de la teoria de la probabilitat, que ens condueix al concepte de mostres aleatòries. En el pla científic, les propietats d'aquesta mena de mostres es basen en els teoremes de la teoria de la probabilitat. En el pla del coneixement comú, la noció que s'ha fet

popular de mostra aleatòria introdueix malentesos i explicacions quasi màgiques. Per aquesta raó, és important popularitzar les idees bàsiques de la lògica del raonament probabilístic.

En el camp de la recerca qualitativa, no hi ha valors numèrics per comparar, i les generalitzacions es fan a través d'identificar algunes classes típiques de processos. Molts investigadors mostren el seu desconeixement científic del concepte de mostra representativa quan demanen que els elements s'escullin a l'atzar.

La predicció probabilística

Crec que és important repassar algunes propietats bàsiques de la teoria de la probabilitat, que constitueixen la lògica de la predicció.

Noció d'independència probabilística. En jugar repetidament a un joc d'atzar, com per exemple llançar una moneda o un dau (però no treure boles d'una urna o loteria sense tornar-les a posar a dins), els resultats que es produiran en les jugades futures són independents dels que s'han produït en les anteriors, en el sentit que estableix la definició següent: les probabilitats que tenen els resultats futurs de sortir no queden modificades pel coneixement dels resultats anteriors. Així, si la primera vegada que s'ha tirat el dau ha sortit un 4, aquest coneixement no canvia la probabilitat dels resultats que poden sortir a la segona tirada. Per aquesta raó, les seqüències de resultats que surten són imprevisibles i desordenades, i no obeeixen a cap llei general.

Distribució de probabilitat i esperança matemàtica. La distribució de probabilitat d'un joc que té un nombre finit de resultats numèrics està formada pel conjunt dels seus resultats possibles i les probabilitats que tenen de sortir. En el cas de llançar un dau perfecte una vegada, la distribució de probabilitat corresponent consta del conjunt de resultats possibles {1,2,3,4,5,6}, cadascun d'ells amb la probabilitat d'1/6. Els gràfics de les distribucions finites de probabilitat sobre el nombre de vegades que pot sortir cara en tirar una moneda moltes vegades –per exemple, diguem-ne 50– són molt útils per entendre algunes lleis de probabilitat i explicar-les, tant als estudiants com als ciutadans. Cadascuna de les distribucions finites de probabilitat té un valor característic, que és l'esperança matemàtica o valor esperat, que es defineix per la fórmula senzilla que podem recordar en el cas del dau perfecte: l'esperança matemàtica és igual a $\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6}$, és a dir, és igual a 3,5. La fórmula general és la suma dels productes de cada valor possible per la seva probabilitat corresponent. Però les expressions *esperança matemàtica* i *valor esperat* requereixen alguns comentaris.

Les lleis empíriques de l'atzar. Al segle XVII, al començament del càlcul de probabilitats, els jugadors van observar dues lleis de regularitat dels resultats estadístics, no individuals sinó agregats. Primera, si tirem una moneda moltes vegades, la freqüència relativa de cares que van sortint s'aproxima a la probabilitat d'aquest resultat. Segona, si tirem un dau moltes vegades, la mitjana de punts per partida que van sortint s'aproxima a l'esperança matemàtica o valor esperat de 3,5. Per aquesta raó, aquest nombre s'anomena l'*esperança matemàtica*. Però veurem que aquesta terminologia pot portar a malentesos si no se n'entén bé el significat estadístic. Les lleis empíriques de l'atzar tenen dos problemes importants que van estimular la recerca matemàtica. D'una banda, no es formulen d'una manera precisa, ja que no indiquen el grau d'aproximació que hi ha entre la probabilitat i les freqüències observades en funció del nombre de repeticions. D'altra banda, aquestes lleis es justifiquen per l'experiència i no són teoremes demostrats. Aquests reptes van portar a formular i demostrar els teoremes que es coneixen com les lleis dels grans nombres.

Freqüències amb què pot sortir un resultat en una sèrie de repeticions. Suposem que volem tirar una moneda perfecta 50 vegades i que ens preguntem quins són els valors del nombre de cares que poden sortir i quines probabilitats hi ha que surtin. El càlcul de probabilitats, que no

desenvoluparé aquí, ens dona la solució a través d'un cas particular de la llei binomial, de la qual presentaré la distribució de probabilitat al gràfic de la figura 1:

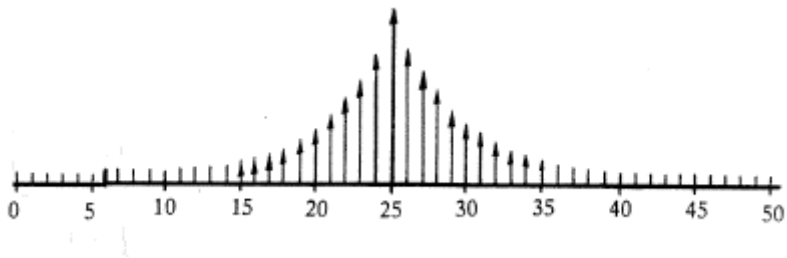


Figura 1

D'aquestes probabilitats, n'indicaré els valors següents, amb dues xifres decimals:

Freqüències:	18	19	20	21	22	23	24	25	...
Probabilitats:	0,01	0,02	0,04	0,05	0,07	0,09	0,10	0,11	...

...	26	27	28	29	30	31	32
	0,10	0,09	0,07	0,05	0,04	0,02	0,01

L'esperança matemàtica d'aquesta distribució de probabilitat és 25 i moltes vegades caiem en la temptació de dir que esperem que surtin 25 cares, però hem de ser molt més meticulosos en el nostre llenguatge ja que, en una sèrie de 50 tirades, la probabilitat que surtin 25 cares és només d'un 11 %. Tanmateix, el gràfic i la taula anteriors ens permeten veure que la probabilitat dels segments curts centrats en el 25 acumulen una probabilitat important i que els segments extrems en tenen molt poca.

Error aleatori. La diferència entre les freqüències que poden sortir i l'esperança matemàtica es denomina *error aleatori*. A partir del gràfic i la taula anteriors, en podem veure molt fàcilment els valors possibles i les probabilitats. Al meu entendre, es tracta d'un concepte clar, però que s'expressa amb una terminologia molt perillosa. Molta gent pensa que havien de sortir 25 cares i que, si no és així, és perquè l'atzar s'ha equivocat. Crec que és molt difícil canviar aquesta expressió, però considero que seria més adequat parlar de "la diferència aleatòria respecte de l'esperança matemàtica".

Predicció i risc. Si ens pregunten si, en tirar una moneda perfecta 50 vegades, poden sortir 50 cares, direm, sense pensar-ho gaire, que en principi és possible que surtin però que, de fet, no sortiran, ja que la probabilitat d'aquest esdeveniment és molt petita (exactament és $(1/2)^{50}$ i, aquest valor és molt més petit que 1 dividit per 1.000 bilions, és a dir, 1 dividit per 1.000 milions de milions. Aquest argument és correcte però la seva formulació en els termes "poden sortir però no sortiran" és una contradicció lògica. Per evitar aquest parany, podríem dir "poden sortir però esperem que no surtin", o anar una mica més lluny i introduir els conceptes següents: l'enunciat "no sortiran 50 cares" és una predicció i la probabilitat o el risc que no es compleixi és $(1/2)^{50}$. En general, les prediccions sobre els resultats que sortiran en els jocs aleatoris tenen sempre un risc de fallar.

Predicció, interval de tolerància i risc. Continuant amb l'exemple de les 50 tirades, fem la predicció que el nombre de cares estarà comprès entre 21 i 29, és a dir, que difereixi de 25 en menys de 5 unitats. L'interval $[21,29]$ o 20 ± 4 s'anomena *interval de tolerància*, en el sentit que indica la diferència que nosaltres admetem sobre el valor esperat. La taula de probabilitat ens permet calcular, simplement per sumes, la probabilitat que aquesta predicció es compleixi, que és del 82 %, i la probabilitat o risc que no es compleixi, que és el valor complementari del 18 %. Així, a cada interval de tolerància, hi associem un risc. També podem procedir d'una altra manera i, fixat un risc α determinat, podem calcular l'interval de tolerància que té un risc més petit o igual a α . Així, si admetem un risc màxim del 5 %, l'interval de tolerància és $[18,32]$, és a dir, 25 ± 7 .

La primera llei dels grans nombres introdueix, sobre els conceptes anteriors, la idea que podem augmentar el nombre de tirades per aconseguir que la freqüència relativa del nombre de cares que sortiran difereixi de la seva probabilitat segons un interval de tolerància determinat i sota un risc de predicció també determinat. El seu enunciat és el següent: per cada interval de tolerància $t > 0$ i cada risc $\alpha > 0$, existeix un nombre de tirades n a partir del qual la predicció que la freqüència relativa de cares diferirà de la seva probabilitat en menys de t té un risc més petit que α . En la teoria de mostres, normalment s'admet una tolerància del 5 % i un risc del 5 %, els quals s'aconsegueixen amb una mostra de 384 persones, que s'arrodoneix a 400. Aquest nombre ha esdevingut quasi màgic en moltes recerques sociològiques.

Mostres aleatòries. Els conceptes i teoremes anteriors ens permeten fonamentar les propietats de les mostres aleatòries i estudiar-ne la representativitat. Si en una població molt gran volem conèixer, amb una certa aproximació i sempre admetent un risc, el percentatge de gent que té una característica o una opinió determinada, en podem treure una mostra petita que sigui aleatòria, és a dir, que doni a tots els elements de la població la mateixa probabilitat de ser escollits.

Estimació estadística

El problema d'aproximar el percentatge dels elements d'una població que tenen una característica a través del percentatge dels que la tenen en una mostra aleatòria equival al problema d'aproximar la probabilitat d'un joc a partir dels resultats observats. Conceptualment, aquest tema és el més senzill de la teoria de l'estimació estadística i està relacionat amb les lleis dels grans nombres. De fet, constitueix el problema invers del de predicció. En efecte, mentre que en la predicció coneixem la probabilitat i volem predir la freqüència que en sortirà, en l'estimació coneixem les freqüències i volem aproximar la probabilitat. No en comentaré els aspectes específics, ja que la lògica ha quedat establerta en el tòpic de la predicció; però n'indicaré alguns abusos i mals usos pràctics.

En el cas d'un estudi sobre la població d'una ciutat gran determinada pels veïns que estan empadronats, en podem treure rigorosament una mostra aleatòria procedint de la manera següent: partim del padró municipal, que ens dona una llista numerada de tots els veïns, amb els seus noms i domicilis. Emplem un programa informàtic que generi nombres aleatoris i amb aquests nombres identifiquem les persones que formaran la mostra.

Normalment, la situació és més difícil ja que no tenim un mètode per identificar i accedir a tota la població per dissenyar la mostra. A més a més, molta gent no respon o no diu la veritat. Les enquestes de molts diaris i altres mitjans de comunicació, que contesten les persones que volen, són un exemple destacat de com no s'ha de fer estadística. No vull pas dir que els resultats no indiquin res, però no tenen cap rigor. Potser tots plegats hauríem de promoure una actitud més científica i no fer passar bou per bèstia grossa. Crec que els periodistes, que estimen la seva contribució a la democràcia, hi col·laborarien amb entusiasme.

Parany conceptuais i de llenguatge

en les aplicacions científiques d'alguns tests estadístics

El Tribunal de la Santa Inquisició feia llistes de les obres que no es podien llegir i ara les comunitats científiques fan llistes indexades de les revistes en les quals hem d'intentar publicar. Moltes revistes que pertanyen a les categories més altes han publicat normalment investigacions basades en tests estadístics, aparentment aplicats d'una manera rigorosa, que han portat a conclusions catastròfiques. L'origen d'aquests desastres és tant el desconeixement de tota la lògica del test, especialment del risc de segona espècie i la funció d'utilitat, com l'ús d'un llenguatge que pot portar a confusions, especialment les expressions *significació estadística* i *diferències significatives*. Els casos típics, en ciències molt diverses, són el resultat de sotmetre a un test estadístic, basat en mostres molt petites, una hipòtesi i concloure que l'evidència empírica confirma que és vertadera, ja que les diferències observades no són significatives.

Per denunciar aquesta situació, Stephen Ziliak i Deirdre McCloskey han publicat el llibre *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives* (University of Michigan Press, 2007). El títol no pot ser més explícit i demolidor. El llibre inclou una presentació conceptual, una indagació sobre les publicacions en diverses ciències i una investigació històrica sobre la formació del llenguatge estadístic. L'expressió *statistical significance*, que porta a molts malentesos, va ser introduïda per Sir Ronald Fisher (1890-1962), un dels pares de l'estadística moderna, no sense ser criticada per altres estadístics com William Gosset. En aquesta presentació em centraré en el test de Student.

Marc conceptual del test de Student. La teoria de les mostres petites va començar amb els treballs de William Gosset, que va publicar amb el pseudònim de "Student". Gosset era un químic i un home de negocis que treballava a la destil·leria de cervesa Guinness de Dublín i estudiava alguns factors que podien influir en la seva qualitat. Aquest problema el va portar a establir el marc conceptual següent: suposem que els resultats que observem d'un fenomen determinat obeeixen a una llei

de probabilitat normal que té una esperança matemàtica m i una variància σ^2 , que no coneixem. En aquest context, volem controlar la hipòtesi H_0 , que anomenem *hipòtesi nul·la*, segons la qual $m=100$, a partir dels resultats d'una mostra petita, de mida $n=16$ observacions. Gosset va calcular la mitjana estadística $\bar{x}$ i la variància estadística S^2 que corresponien als valors observats. A partir d'ells va introduir la fórmula del test i les distribucions de probabilitat de Student.

Simplificació del problema. Per simplificar el problema de manera que la seva lògica sigui més entenedora, suposaré que coneixem la variància teòrica σ^2 per investigacions anteriors (a la pràctica, aquest supòsit no es dona quasi mai) i que el seu valor és $\sigma^2 = 2$.

Llei de probabilitat de cada valor observat i de la mitjana. Tal com indica la figura 2, com que cada resultat observat ha sortit d'acord amb una llei normal d'esperança matemàtica m i de variància $\sigma^2 = 2$ (és a dir, $\sigma = \sqrt{2}$), la mitjana $\bar{X}$ dels setze valors observats ha sortit d'acord amb una llei

normal d'esperança m i de variància $\sigma'^2 = \frac{\sigma^2}{16} = \frac{2}{16}$ (és a dir, $\sigma' = \frac{\sqrt{2}}{4} = 0,35$).

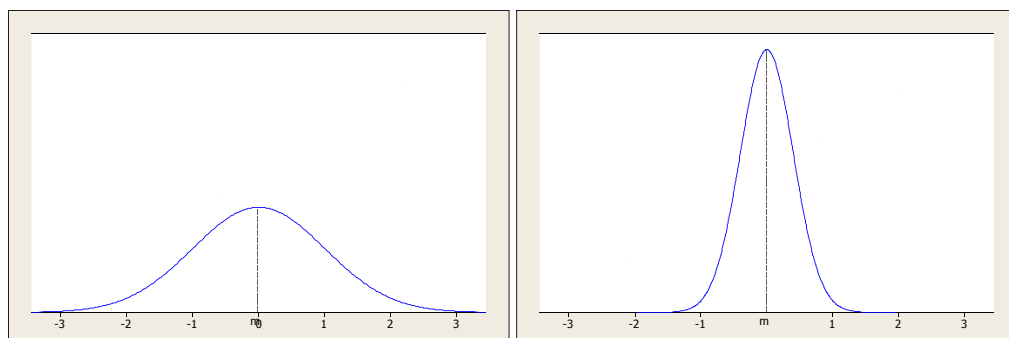


Figura 2

Problema previ: predicció del valor que pot tenir la mitjana $\bar{X}$, si suposem que l'esperança matemàtica és $m=100$. En aquest problema previ, ignorem de moment el valor que hem calculat de la mitjana estadística $\bar{X}$ i suposem que l'esperança matemàtica és efectivament $m=100$. En aquest supòsit, poden preveure l'interval de tolerància en el qual sortirà la mitjana $\bar{X}$. Si admetem un risc de predicció $\alpha=5\%$, l'interval de tolerància, calculat a través de les taules de la llei normal, és $t = 1,96 \sigma' \approx 2\sigma' = 2 \times 0,35 = 0,70$. És a dir, l'interval de tolerància és $100 \pm 0,70$, o en altres termes $[99,3 - 100,70]$.

Transformació del problema de predicció en el problema del test estadístic: risc de primera espècie i regla de decisió. L'interval de tolerància que hem calculat serà molt important per establir la regla de decisió del test estadístic, pel raonament següent: en sotmetre a un test estadístic la hipòtesi nul·la $H_0: m=100$, partim de dues idees intuïtives molt importants: en primer lloc, admetem que la hipòtesi nul·la és vertadera mentre no hi hagi una evidència empírica en contra. En segon lloc, si aquesta hipòtesi és vertadera, la mitjana estadística $\bar{X}$ que en sortirà estarà més o menys a prop de $m=100$, i si és falsa s'allunyarà d'aquest valor. Com que, si H_0 és vertadera, sabem que, amb un risc del 5%, $\bar{X}$ sortirà en l'interval $[99,30 - 100,70]$, podem introduir els conceptes següents: amb un risc de primera espècie $\alpha=5\%$, la regla de decisió queda determinada de la manera següent: si la mitjana estadística $\bar{X}$ està compresa entre $[99,30 - 100,70]$, acceptem la hipòtesi nul·la $m=100$ (zona d'acceptació) i, si queda fora d'aquest interval (zona de rebuig), considerem que la hipòtesi és falsa. El risc de primera espècie té el significat següent: en cas que la hipòtesi $m=100$ sigui vertadera, la probabilitat que la considerem falsa és $\alpha=5\%$. Quan la diferència entre 100 i la mitjana estadística és menor que 0,7 diem, en el llenguatge de Fisher, que no és significativa i, quan és més gran, diem que sí que és significativa.

En el test de Student seguim aquestes idees, però com que no en coneixem la variància teòrica σ^2 , la substituïm per la variància estadística s^2 , que hem calculat. Aquest canvi ens porta a passar de les lleis normals a les lleis de Student, que també estan tabulades i condueixen a intervals més llargs. Gosset també va estudiar el problema de sotmetre a un test si dues poblacions normals (o una població en períodes diferents) tenen la mateixa esperança matemàtica o si són diferents. Un gran nombre de treballs de recerca es desenvolupen aplicant aquests tests, que, en principi, són correctes. El problema és que els investigadors només s'ocupen, o només coneixen, aquesta part del test i n'ignoren altres aspectes bàsics. Al cap i a la fi, hem dissenyat un test admetent que la hipòtesi nul·la $H_0: m=100$ és vertadera, quan el que ens preocupa és que la hipòtesi sigui falsa.

Hipòtesi alternativa i risc de segona espècie. Suposem que ens fixem en el cas que l'esperança matemàtica sigui $m=100,5$ i, paral·lelament, en el cas que sigui $m=99,5$ perquè poden tenir conseqüències molt importants en el nostre tema de recerca. Així, direm que $m=100,5$ i $m=99,5$ constitueix la hipòtesi alternativa H_A i calcularem la probabilitat que, si m és realment igual a $100,5$, el test no ho detecti i accepti que m és igual a 100 . D'acord amb l'exemple simplificat de $\sigma^2=2$, podem fer el raonament següent: tal com indica la figura 3, en la qual hem amplificat l'escala horitzontal, la zona d'acceptació del test és l'interval $[99,30 - 100,70]$, però ara la llei normal no està centrada sobre 100 , com la de l'esquerra del dibuix, sinó sobre $100,5$, com la de la dreta.

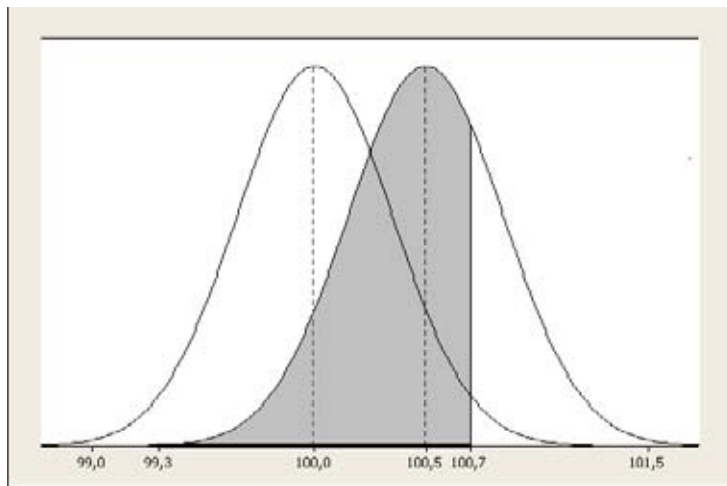


Figura 3

Ara bé, fent només una ullada al gràfic, ens adonem que la probabilitat que busquem, és a dir, la probabilitat que, suposant que la hipòtesi H_0 sigui falsa i m sigui $100,5$, passi el test com a vertadera (hem ratllat aquesta probabilitat) és més gran que $\frac{1}{2}$. Si la calculem, emprant les taules de la llei normal obtenim que el seu valor és del 69 %. Recordem que el seu sentit és el següent: en cas que la hipòtesi nul·la $H_0: m=100$ sigui falsa i la hipòtesi alternativa $H_A: m=100,5$ sigui vertadera, la probabilitat que el test accepti H_0 com a vertadera, que s'anomena *risc de segona espècie* i es representa per β , és del 69 %. Així, aquest test té una probabilitat molt elevada que, si la hipòtesi H_0 és falsa, passi com a vertadera.

Mentre que el camp de l'enginyeria i del control de la qualitat té una tradició molt forta que destaca el risc de segona espècie, les ciències socials moltes vegades l'ignoren i una gran part de les investigacions que publiquen a partir de mostres petites afirmen que la hipòtesi és confirmada empíricament. Aquest problema s'estén més enllà del test de Student i afecta els tests sobre els coeficients de correlació i de regressió, que moltes vegades arriben a la conclusió que els resultats empírics observats no difereixen significativament de 0.

En aquest estudi, és important destacar de quina manera alguns termes estadístics poden portar –i, de fet, han portat– a interpretacions incorrectes. En primer lloc, de la mateixa manera que l'expressió *error aleatori* no es refereix a cap error real sinó a la diferència entre l'esperança matemàtica i el valor que pot sortir, tampoc l'expressió *error estàndard* no indica cap error sinó que denota la dispersió típica de la llei de probabilitat i, d'alguna manera, mesura la variabilitat dels resultats. En canvi, en el test estadístic sí que hi ha dos possibles errors reals: l'error de concloure

que la hipòtesi nul·la és falsa quan realment és vertadera (que porta el risc de primera espècie) i l'error de considerar que la hipòtesi nul·la és vertadera quan realment és falsa (que porta el risc de segona espècie). Aquest segon error és el que molts treballs no tenen en compte. En segon lloc, l'acceptació o no-acceptació de la hipòtesi nul·la depèn de l'error estàndard, és a dir, de la variabilitat dels resultats i no té res a veure amb el significat científic i pràctic del tema. És a dir –i aquesta n'és la crítica més important–, la significació estadística és un concepte relacionat amb la variabilitat i no té res a veure amb la importància científica dels valors esperats. Molts investigadors es pensen que el concepte clau del test és la significació estadística i ignoren la importància científica dels valors teòrics.

Una objecció que es podria fer a l'exemple que hem presentat és que hi hem introduït unes dades amb les quals el risc de segona espècie és molt alt. El fet és que en molts problemes reals la variabilitat molt gran dels resultats estadístics observats i la importància de les variacions petites de les esperances matemàtiques fan que els tests estadístics normalment concloguin que “les diferències no són significatives” o que “no hi ha cap evidència científica d'un canvi”, sense preocupar-se del risc de segona espècie, que és molt important. Crec que aquest és un aspecte dels debats de fa uns quants anys sobre el canvi climàtic i l'escalfament de la terra. Em sembla que alguns científics, encara que anéssim a les calderes d'en Pere Botero (no vull dir que s'hi haguessin de quedar), diríem que no tenien cap evidència científica de l'escalfament. Amb enfocaments més documentats, el llibre *El culte a la significació estadística* presenta molts exemples d'investigacions en diverses ciències. Per exemple, mostra com els tests estadístics han portat a considerar esquizofrèncics molts malalts que no ho són.

Els llenguatges estadístics i la llengua catalana

Dolors Presagué Vila
Estadística de l'Idescat

1. Els llenguatges estadístics en l'estadística oficial
2. Aspectes metodològics de les classificacions estadístiques
3. La terminologia estadística i les classificacions estadístiques. La col·laboració amb el TERMCAT
4. Conclusions i reflexions generals

Abstract

L'objectiu de l'article se centra a explicar la metodologia de treball que aplica l'Idescat a l'hora de dur a terme les classificacions estadístiques (llenguatges estadístics), que ens permeten comparar els resultats estadístics amb els d'altres àmbits temporals i espacials en el marc de l'estadística oficial, assenyalant la llengua catalana com a eina instrumental bàsica per elaborar-les.

La metodologia de treball a l'entorn de les classificacions estadístiques s'ha anat consolidant amb la seva pràctica i es concreta en tres processos molt importants, que són el procés de normalització a la llengua catalana de la classificació estadística, el procés de socialització de la classificació estadística als agents del Sistema estadístic de Catalunya i, finalment, el procés d'oficialització que dona com a resultat una classificació estadística normativitzada, és a dir, de compliment obligatori.

En l'execució d'aquesta metodologia de treball, s'explicarà la importància del model de llengua que s'aplica en el camp de les classificacions estadístiques, model de llengua que s'ha anat consolidant des de la Unitat de llengua i estadística de l'Idescat amb la col·laboració del TERMCAT, per tal d'aconseguir com a objectiu prioritari la qualitat i la coherència d'aquests llenguatges estadístics, que, alhora, enriqueixen i fan créixer el llegat terminològic de la nostra llengua en una societat tan globalitzada i tan especialitzada com és l'actual.

1. Els llenguatges estadístics en l'estadística oficial

El fenomen de la globalització en el qual està immersa la societat actual provoca i condueix a intercanvis entre llengües i sectors temàtics especialitzats que donen lloc a un creixement social important de la llengua i de la ciència, multiplicant-se cada vegada més els llenguatges especialitzats.

Totes les branques del coneixement en general i, en aquest cas concret les ciències, necessiten disposar de mots especialitzats que puguin designar els seus conceptes. Aquests mots, utilitzats amb un significat unívoc en una ciència o una disciplina determinada per tal d'evitar una correspondència equívoca entre els conceptes i la seva expressió, reben el nom de **termes**, i tal com apunta la seva definició són elements imprescindibles a l'hora de poder expressar o descriure qualsevol concepte de forma inequívoca.

En el món actual, en què els avanços tecnològics progressen a un ritme vertiginós, i per tant, també a un ritme vertiginós hem de descriure i expressar nous conceptes, la terminologia, entesa com la disciplina que s'encarrega de l'estudi i de la recopilació dels termes especialitzats, esdevé cada vegada més una eina indispensable per a la investigació en general, i sens dubte també per a l'estadística, entesa com la ciència que recull, classifica, analitza i interpreta les dades relatives a un conjunt d'elements.

L'estadística, com qualsevol altra ciència, necessita expressar-se per mitjà d'un llenguatge precís i adequat a les seves necessitats, per tant, les necessitats terminològiques de l'estadística són tan àmplies com ampli sigui l'univers que es vol estudiar. En aquest sentit, és molt important assenyalar que la terminologia és imprescindible en el camp de les **nomenclatures i classificacions estadístiques** que es concreten en els **llenguatges estadístics** que serveixen per presentar les dades estadístiques de forma coherent, ordenada i susceptibles de ser comparades amb altres dades d'altres àmbits espacials i temporals.

Atesa la importància i la interdependència d'ambdues disciplines científiques, llengua i estadística, és òbvia i s'intueix, doncs, fàcilment la rellevància que té l'aplicació d'una metodologia terminològica en el camp de l'estadística. L'Institut d'Estadística de Catalunya (Idescat), en aquest sentit, i arran de la feina feta i els esforços esmerçats en el camp dels llenguatges estadístics, ha aconseguit establir un model de llengua de les classificacions estadístiques dins l'àmbit de l'activitat estadística oficial.

Les nomenclatures i classificacions estadístiques constitueixen un sistema bàsic de terminologia que garanteix la recollida, la presentació, l'anàlisi i la comparabilitat de les dades estadístiques en l'espai i el temps, i s'inclouen en un concepte ampli i innovador de les **activitats estadístiques oficials** com és el de les **activitats estadístiques instrumentals**, concepte que recull de forma expressa la Llei 23/1998, de 30 de desembre, d'estadística de Catalunya, que té per objecte regular l'estadística d'interès de la Generalitat en exercici de la competència exclusiva que li confereix l'Estatut d'autonomia de Catalunya, i també les activitats per dur-la a terme.

Als efectes d'aquesta Llei s'entén per activitats estadístiques instrumentals aquelles activitats estadístiques prèvies o complementàries a les activitats estadístiques que condueixen a l'obtenció, la recopilació, l'elaboració i l'ordenació sistemàtica de les dades, i a la conservació, l'emmagatzematge, la publicació i la difusió de resultats. Són activitats legalment exigibles i tècnicament necessàries per poder complir els requisits que estableix la legislació sobre estadística, i en són exemples les activitats estadístiques instrumentals de recerca i desenvolupament tècnic, metodològic i normatiu en el camp estadístic i, és justament, dins el camp de la metodologia on cal incloure els llenguatges estadístics, és a dir, les nomenclatures i classificacions estadístiques.

De fet, la mateixa Llei en l'apartat de les funcions assignades a l'Idescat estableix, entre altres, en

l'apartat q de l'article 10 la funció d'**adaptar, normalitzar, revisar o establir les classificacions estadístiques**, les definicions i els codis, les referències territorials i els trams de desagregació estandarditzats, i elaborar les propostes normatives pertinents, tot salvant la compatibilitat i la correspondència amb els homòlegs d'ús obligatori en els àmbits estatal, europeu i internacional.

En aquest sentit, per tant, queda palesa la vinculació i la relació entre la llengua i l'elaboració de les classificacions estadístiques en el sentit que la llengua i la terminologia constitueixen instruments del tot necessaris a l'hora d'elaborar aquests instruments estadístics que són previs i complementaris a les activitats estadístiques de producció i difusió. La importància de la producció i la difusió de les dades estadístiques radica en el fet que es puguin comparar amb altres dades d'altres àmbits territorials o temporals, i l'ús dels llenguatges estadístics garanteix la seva comparabilitat, ja que de res serviria produir i difondre dades estadístiques de Catalunya, per exemple, si no es poguessin comparar amb dades d'altres àmbits territorials, ja que s'hauria dut a terme una feina o un estudi *ad hoc* sense un valor afegit important com és el d'obtenir resultats interpretables des de l'òptica conceptual d'integració dins el sistema.

En aquest context normatiu, els llenguatges estadístics han de satisfer unes determinades condicions: respondre a les necessitats dels principals productors i usuaris de la informació estadística, ser senzills i precisos, i avançar-se pel que fa a futures necessitats d'ampliació del llenguatge derivades de l'evolució cap a noves tecnologies i productes.

Aquest conjunt de condicions que han de satisfer els llenguatges estadístics estandarditzats són tan rellevants i imprescindibles en el camp de l'estadística oficial que s'han inclòs en el **Codi de les bones pràctiques de les estadístiques europees**,¹ tot i que d'una forma molt implícita, perquè malgrat la importància de la llengua, concebuda com una eina bàsica a l'hora d'elaborar els llenguatges estadístics estandarditzats, tan sols apareix entre línies i de forma implícita com un indicador del principi de l'accessibilitat i claredat de les estadístiques oficials, i ni tan sols s'apunta, ni que sigui de forma nítida, com un indicador del principi de qualitat de les estadístiques oficials.

Per acabar aquest apartat, es pot concloure que els llenguatges estadístics són instruments estadístics necessaris que per a la seva elaboració és totalment imprescindible fer ús de la terminologia precisa que demana cadascun dels camps temàtics que es vol estudiar, és a dir, es requeriran molts termes i molt diferents en funció de si s'estan recollint i tractant dades estadístiques referides bé a l'activitat econòmica que duen a terme les empreses, bé als productes que són objecte de comerç amb l'estranger, bé a l'ocupació i al nivell de formació dels ciutadans o bé a altres variables susceptibles d'estudi i de tractament estadístic o bé d'aprofitament administratiu.

La interdependència entre la llengua i l'estadística en el camp de les nomenclatures i classificacions estadístiques és clara i manifesta i, en aquest sentit l'Idescat treballa de forma conjunta amb els especialistes de la llengua per tal d'anar avançant en la consecució de la qualitat global dels termes estadístics amb l'objectiu d'intentar arribar una mica més lluny i fer arribar a les autoritats de l'estadística oficial la necessària incorporació del bon ús de la llengua com un indicador del principi de qualitat en el Codi de les bones pràctiques estadístiques.

1 Codi de les bones pràctiques de les estadístiques europees. Adoptat pel Comitè del Programa estadístic del 24 de febrer del 2005 i publicat en la Recomanació de la Comissió, de 25 de maig de 2005, sobre la independència i la responsabilitat de les autoritats estadístiques nacionals i comunitàries.

2. Aspectes metodològics de les classificacions estadístiques

Metodologia de treball pròpia

L'objectiu d'aquest apartat se centra a explicar la metodologia de treball associada a l'elaboració de les classificacions estadístiques que l'Idescat ha anat creant i consensuant mitjançant la seva pràctica en el marc del Sistema estadístic de Catalunya (SEC) i, que, actualment, es caracteritza pel seu elevat grau de consolidació.

De fet, el que es pretén explicar és que en virtut de les funcions que la Llei 23/1998, de 30 de desembre, d'estadística de Catalunya, assigna a l'Idescat en matèria de nomenclatures i classificacions estadístiques (adaptar, normalitzar, revisar o establir les classificacions estadístiques.....), l'Idescat ha desenvolupat una **metodologia de treball pròpia** que es concreta en tres processos molt importants. El primer és el **procés de normalització** a la llengua catalana de la classificació estadística oficial que es vol adaptar, el segon és el **procés de socialització** de la classificació estadística normalitzada als agents del Sistema estadístic de Catalunya, i, finalment, el tercer **procés consisteix en l'oficialització** de la classificació estadística, que és el resultat una norma de compliment obligatori.

En aquest context, i per tal d'entendre la raó de ser d'aquests tres processos detallats anteriorment, es fa necessari definir i diferenciar els conceptes d'harmonització estadística i normalització estadística, que, al mateix temps, s'expliquen a partir del concepte de la integració. La integració es defineix com la relació i coordinació que s'estableix dintre i entre els diferents sistemes estadístics, amb l'objectiu d'assolir la convergència i la comparabilitat en l'àmbit internacional, comunitari i estatal de la informació estadística, a banda d'assolir altres objectius molt importants com ara, en el cas del Sistema estadístic de Catalunya, l'avanç en la normalització de la terminologia catalana i l'aplicació d'un **model de llengua propi** en l'àmbit de l'estadística oficial.

És important saber que l'harmonització estadística i la normalització estadística conflueixen en l'objectiu de comparar la informació estadística, i tan sols difereixen en el seu abast; és a dir, el concepte d'harmonització permet definir un procés, un conjunt d'actuacions, on es troben immersos molts organismes estadístics oficials, mentre que, el concepte de normalització permet definir una actuació concreta d'aquell procés i serà duta a terme per un d'aquests organismes estadístics oficials amb l'objectiu comú de comparar les diferents fonts d'informació.

El procés d'homogeneïtzació internacional en l'àmbit de les classificacions estadístiques està liderat i coordinat per la Divisió Estadística de les Nacions Unides (UNSD) i per l'Oficina Estadística de les Comunitats Europees (Eurostat). Per exemple, l'elaboració de la *Classificació industrial internacional uniforme per a totes les activitats econòmiques Revisió 4 (CIIU-Rev.4)* i la *Nomenclatura estadística d'activitats econòmiques a la Comunitat Europea Revisió 2 (NACE-Rev.2)*, liderades per Nacions Unides i Eurostat, respectivament, constituïran els models harmonitzats i integrats de partida amb la finalitat que els organismes estadístics oficials dels estats membres de la Unió Europea duguin a terme els processos de normalització concrets i elaborin les seves versions nacionals de les classificacions d'activitats econòmiques. Un exemple de normalització estadística és la recent *Classificació catalana d'activitats econòmiques 2009 (CCAE-2009)*, elaborada per l'Idescat, a partir de la *Clasificación nacional de actividades económicas 2009 (CNAE-2009)*.

També val la pena assenyalar que existeixen diferències conceptuals entre els termes nomenclatura i classificació que són utilitzats moltes vegades com a sinònims, mentre que, per definició, corresponen a conceptes substancialment diferents.

El terme *nomenclatura* es defineix com una llista objectiva de terminologia de les característiques

que presenta un fenomen en estudi i, perquè aquesta llista sigui manejable, les característiques s'agrupen en categories en funció de propietats comunes amb la qual cosa s'entra en el camp de les classificacions.

En aquest sentit, el concepte de *classificació* implica l'organització i l'ordenació dels ítems d'una nomenclatura seguint uns criteris subjectius de classificació que actuen com a directrius per determinar la seva estructura, dins la qual s'estableixen diferents nivells d'agregació.

Les classificacions estadístiques s'agrupen segons **tipologies temàtiques anomenades “famílies”**, i les principals són les **classificacions econòmiques** (activitats econòmiques i productes) i les **classificacions demogràfiques i socials** (ocupacions, educació i sanitàries) i ens permetran avaluar els fenòmens de la realitat econòmica i social.

Una vegada definits els conceptes nuclears a l'entorn de les nomenclatures i classificacions estadístiques, s'explicarà la metodologia de treball de l'Idescat, tot detallant-se cadascun dels processos apuntats.

Procés de normalització

El procés de normalització de la classificació estadística a la llengua catalana es duu a terme una vegada s'ha elaborat i aprovat la classificació estadística d'àmbit internacional i comunitària, atès que es tracta d'una actuació de normalització concreta fruit d'un procés global d'harmonització, la finalitat del qual és aconseguir un sistema integrat de llenguatges estadístics estandarditzats.

En aquest context, on el principi d'integració és bàsic a l'hora de treballar, les actuacions de l'Idescat, com a gestor del Sistema estadístic de Catalunya, cal entendre-les a partir dels resultats de les actuacions de l'Institut Nacional d'Estadística (INE), que gestiona el Sistema estadístic estatal.

L'INE respon, al mateix temps, a les actuacions de l'Oficina Estadística de les Comunitats Europees (Eurostat), organisme que gestiona el Sistema estadístic europeu (SEE) i que alhora trasllada les seves propostes d'actuació provinents dels organismes estadístics dels estats membres a la Divisió Estadística de les Nacions Unides perquè les tingui en consideració en les seves recomanacions d'àmbit internacional, juntament amb d'altres provinents d'altres sistemes estadístics.

En aquest procés de treball caracteritzat per la integració, els **grups de treball** constitueixen els mecanismes o eines de treball (fòrums de cooperació) que s'estableixen entre Eurostat i les oficines estadístiques nacionals dels estats membres, de les quals l'Idescat en forma part amb l'objectiu de traslladar-hi les consideracions formulades pels agents del Sistema estadístic de Catalunya a l'efecte.

El procés de normalització de la classificació estadística que duu a terme l'Idescat consisteix en la traducció i adaptació a la llengua catalana de la versió castellana que elabora l'INE. La versió castellana de la classificació es pot trobar en fase d'elaboració, és a dir, en un estadi que va avançant de forma paral·lela a l'esborrany de la classificació comunitària, o bé ja aprovada per una norma estatal que també, i, òbviament estarà supeditada a la norma comunitària.

El procés de traducció a la llengua catalana es fa a partir d'uns criteris lingüístics que l'equip de planificadors lingüístics de l'Idescat facilita i recomana als nomenclaturistes. La finalitat d'aquests criteris consisteix a donar una pauta a l'hora d'elaborar el primer esborrany de l'estructura i les notes explicatives de la classificació, tot detallant i assessorant, per exemple, pel que fa a l'ús dels articles o preposicions en les enumeracions dels diferents ítems, bé pel que fa a l'ús d'expressions innecessàries, bé pel que fa a la supressió de possessius innecessaris, bé pel que fa a la precisió i claredat en les expressions de quantitats, entre altres.

Una vegada es compta amb un esborrany de l'estructura i les notes explicatives de la classificació normalitzada es trasllada als planificadors lingüístics, perquè en facin la correcció lingüística i terminològica en funció dels criteris lingüístics i terminològics que tenen definits, i que revisen constantment d'acord amb els principis d'homogeneïtzació estadística a partir de les versions de la classificació en altres llengües comunitàries.

Efectuada aquesta primera revisió lingüística, l'equip de planificadors lingüístics de l'Idescat trameta al Centre de Terminologia Catalana (TERMCAT) la llista de dubtes terminològics que no els ha estat possible resoldre, perquè els especialistes de la terminologia els donin una solució, ja sigui mitjançant la normalització de neologismes per part del Consell Supervisor, ja sigui trobant la solució a partir de la recerca de formes de termes en els àmbits d'actuació especialitzats, o bé a partir d'altres formes de recerca de les quals en són l'organisme competent.

El TERMCAT actua com a organisme col·laborador de l'Idescat en matèria de nomenclatures i classificacions estadístiques oficials en els programes anuals d'actuació estadística (PAAEs) que despleguen la Llei del Pla estadístic. En definitiva, amb la seva col·laboració es posa de manifest la importància instrumental de la llengua catalana en el procés d'homogeneïtzació dels llenguatges estadístics estandarditzats, que són alhora instruments estadístics necessaris en la producció i difusió de les dades estadístiques oficials.

És important remarcar, per tant, que en el decurs del procés de normalització de la classificació s'hi aplica un **model de llengua** propi i definit per l'equip de planificadors lingüístics de l'Idescat amb la col·laboració del TERMCAT. D'aquesta manera s'estableix un flux de treball entre ambdós organismes ordenat, planificat i molt conscient de la seva importància.

Aquest circuit de treball dissenyat entre l'Idescat i el TERMCAT ha propiciat l'èxit de la integració i l'aplicació d'un model de llengua en els llenguatges estadístics com una bona pràctica estadística a seguir, i que avui l'Idescat té consolidada a través de la seva continuada aplicació a l'hora d'elaborar-los. Justament l'ús d'aquesta bona pràctica estadística i els resultats que se'n deriven emparen i avalen el fet que tant estadístics com terminòlegs estiguem d'acord a incloure de forma explícita la llengua com un indicador del principi de qualitat en el Codi de bones pràctiques estadístiques.

Procés de socialització

Una vegada el TERMCAT compta amb la llista terminològica resolta, la trameta de nou als lingüistes de l'Idescat per tal que puguin presentar als nomenclaturistes una primera proposta de l'estructura de la classificació revisada lingüísticament i terminològicament, i, que, per tant, estigui en condicions de socialitzar-se als agents del Sistema estadístic de Catalunya, i s'inicia així el procés de socialització de la classificació.

De fet, la tendència de l'Idescat en el marc del Sistema estadístic de Catalunya és treballar de forma paral·lela a l'INE, és a dir, el procés d'elaboració i normalització de la classificació catalana es duu a terme de forma paral·lela al procés d'elaboració de la classificació de l'INE. Aquest paral·lelisme es justifica pel fet que en el grup de treball d'àmbit estatal sobre classificacions estadístiques, format per l'INE, representants de les comunitats autònomes, ministeris i el Banc d'Espanya, l'Idescat hi participa i col·labora de forma molt activa.

La participació de l'Idescat en aquest grup de treball d'àmbit estatal es tradueix en l'aportació, sempre que escaigui, de les consideracions i observacions dels diferents agents del Sistema estadístic de Catalunya a la proposta de classificació normalitzada, i que són fruit del procés de socialització i consulta d'aquesta mateixa dins el sistema.

Aquest procés de socialització del llenguatge estadístic té una importància cabdal, perquè, d'una banda, ens permet comunicar-nos amb els organismes especialistes i competents de les diferents matèries que llista de forma ordenada la proposta de classificació; i, d'altra banda, i com a conseqüència d'aquesta consulta als especialistes temàtics, les aportacions terminològiques són nombroses i vitals de cara a l'obtenció d'un llenguatge estàndard que haurà de permetre que les diferents unitats de l'univers objecte d'estudi s'hi puguin identificar i classificar de forma clara, precisa i sense ambigüitats.²

Els agents del Sistema estadístic de Catalunya estan especialitzats en una sèrie de competències específiques (agricultura, indústria, comerç, educació, medi ambient, salut, innovació, tecnologia i comunicacions (TIC), entre altres) que, òbviament, els converteix en els organismes especialistes de la gestió d'aquests camps temàtics, tant des del punt de vista de continguts i coneixements com des del punt de vista terminològic, en el sentit que la mateixa especialització temàtica els requereix cercar termes específics per designar aquells coneixements sectorials que treballen i gestionen diàriament.

En definitiva, l'objectiu del procés de socialització de la classificació rau en el fet que s'acabi elaborant un llenguatge estadístic estàndard de qualitat que integri el codi i la descripció en un "corpus" unívoc que ens permeti llegir, interpretar i comparar la nostra informació estadística d'un àmbit temàtic determinat amb la d'altres sistemes estadístics presents en la societat global.

Per tal d'exemplificar el resultat del procés de socialització de la proposta de classificació dins el Sistema estadístic de Catalunya i la seva importància des del vessant terminològic, a banda de les aportacions de contingut temàtic, s'exposen les consideracions terminològiques del Departament d'Acció Social i Ciutadania i del Departament de Medi Ambient i Habitatge a la proposta de *Classificació catalana d'activitats econòmiques 2009 (CCA-2009)*, que és la darrera classificació estadística elaborada per l'Idescat i aprovada mitjançant el Decret 137/2008, de 8 de juliol, del Govern de la Generalitat de Catalunya.

Aportacions terminològiques del **Departament d'Acció Social i Ciutadania** a l'estructura i a les notes explicatives de la CCA-2009:

En l'estructura de la CCA-2009 les activitats relacionades amb els serveis socials es classifiquen en les divisions (2 dígits) 87 i 88 de la classificació, i, en el codi 8891 *Activitats d'atenció diürna de menors* es proposa canviar el terme **menors** pel terme **infants**. Per tant, en el Decret de la CCA-2009 aquest codi 8891 apareix amb la descripció oficial *Activitats d'atenció diürna d'infants*.

La denominació **infants** és l'expressió d'ús comú en l'àmbit dels serveis socials, i, en canvi, la denominació **menors** és més pròpia de l'àmbit de la justícia juvenil.

Quant a les notes explicatives dels codis de la CCA-2009 referits als serveis socials, s'han fet també diverses propostes terminològiques que s'han inclòs en la redacció definitiva de les notes explicatives. Es tracta de les substitucions terminològiques següents:

- **orfenats** per **centres residencials d'acció educativa**
- **centre d'acolliment temporal per a persones sense llar** per **residència d'estada limitada**
- **internats i llars infantils** per **centres d'acolliment**
- **centres de convalsència** per **centres sociosanitaris**

2 És important assenyalar que hi ha classificacions estadístiques molt específiques, com ara la Nomenclatura Combinada, classificació de productes que es fa servir per tabular les dades de comerç amb l'estranger, que no se sotmet als processos de socialització i oficialització estàndards donat el seu volum i al fet que varia cada any. En aquest cas, l'Idescat rep l'assessorament habitual del TERMAT i la col·laboració dels serveis lingüístics dels ports de Barcelona, Tarragona i de les Illes Balears.

- **comunitats de jubilats amb atenció continuada** per **habitatges tutelats per a gent gran**
- **establiments per a persones assistides** per **establiments per a persones assistides: centres residencials i habitatges amb serveis comuns per a persones amb discapacitat**

Aportacions terminològiques del **Departament de Medi Ambient i Habitatge** a l'estructura de la CCAE-2009:

Les activitats relacionades amb la silvicultura es classifiquen en la divisió (2 dígits) 02 de la classificació, i, en el *codi 022 Explotació de la fusta* es proposa canviar el terme **explotació** pel terme **aprofitament**. Per tant, en el Decret de la CCAE-2009 aquest codi 022 apareix amb la descripció oficial **Aprofitament de la fusta**.

El valor afegit des del vessant terminològic que es deriva del procés de socialització del llenguatge estadístic als agents competents i especialistes de la gestió de cada secció econòmica específica és obvi, perquè el resultat és un increment extraordinari de la qualitat de la llengua al llarg de tota la classificació en el sentit de precisió, claredat i imparcialitat.

El procés de socialització constitueix una garantia de l'ús del llenguatge estadístic elaborat a partir del model de llengua definit prèviament perquè inclou la terminologia específica que han aportat tots els sectors econòmics o socials representats pels seus interlocutors tècnics o col·laboradors.

Reprement el fil de la metodologia de treball de l'Idescat, explicar que una vegada finalitzat el procés de socialització, la proposta de classificació es trameta de nou a l'equip de planificadors lingüístics i, si escau al TERMCAT, perquè en faci la revisió i validi lingüísticament les aportacions lliurades pels diferents agents del sistema. I, al mateix temps, aquestes aportacions es trameten al grup de treball d'àmbit estatal que lidera l'INE amb la finalitat que tots els membres del grup les avaluin i, sempre que escaigui, s'incorporin també a la versió castellana de la classificació.

Val a dir que són nombroses les aportacions del Sistema estadístic de Catalunya que s'han incorporat a les classificacions d'àmbit estatal, tant de contingut temàtic com de forma, i que, per tant, també en el Sistema estadístic estatal queda palès i s'ha fet ressò, d'una banda, del model de llengua aplicat als llenguatges estadístics, i, d'altra banda, de la nostra dedicació per assolir la màxima qualitat de la llengua utilitzada en els llenguatges estadístics.

Procés d'oficialització

Arribat a aquest punt, tan sols resta l'última fase del procés que ens conduirà a l'oficialització de la proposta de la classificació, i que consisteix a presentar l'estructura de la classificació i els elements que la complementen a la Junta de Govern de l'Idescat (un òrgan col·legiat que té atribuïts dos tipus de funcions: les funcions orientades a la gestió i a la coordinació del sistema estadístic de Catalunya i les funcions que li corresponen com a òrgan de direcció de l'Idescat) i al Consell Català d'Estadística (òrgan de caràcter **consultiu i d'assessorament** de l'Idescat. Hi estan representats diferents sectors socials amb l'objectiu de recollir les seves opinions sobre les necessitats d'informació estadística i verificar si s'ajusten a les demandes de la societat), perquè li donin la seva aprovació.

A partir d'aquest moment, s'inicia la preparació de tots els documents necessaris, així com les gestions necessàries amb el Departament d'Economia i Finances, amb l'objectiu que la classificació es normativitzi i esdevingui una classificació estadística oficialitzada, i, en conseqüència, de compliment obligatori en l'àmbit de l'estadística oficial.

En qualsevol cas, és important tenir present que, tot i el paral·lelisme que es dona a l'hora de dur a

terme el complex procés d'elaboració de la classificació catalana *versus* la classificació espanyola, sempre s'oficialitzarà després de fer-ho aquesta darrera, i aquesta, a la vegada, s'oficialitzarà amb posterioritat a la comunitària.³

3. La terminologia estadística i les classificacions estadístiques

La col·laboració amb el TERMCAT

La importància que té l'aplicació rigorosa d'una metodologia terminològica és òbvia en els llenguatges estadístics, i, per tant, és lògic que el TERMCAT figure com a organisme col·laborador a l'hora de dur a terme les actuacions estadístiques relacionades amb l'activitat estadística Nomenclatures i classificacions estadístiques oficials, en el decurs del desplegament dels programes anuals d'actuació estadística (PAAE) que donen a resposta al Pla estadístic de Catalunya (PEC) vigent.

Cal esmentar que la denominació de les categories es farà mitjançant termes, la qual cosa requerirà una precisió molt gran per tal que tant l'abast dels conceptes continents com el dels continguts quedi perfectament delimitat i fixat. És molt important evitar la sinonímia (relació d'identitat semàntica entre mots) tant entre termes d'una mateixa categoria jeràrquica com de categories jeràrquiques diferents. També caldrà resoldre els casos d'homonímia (identitat fònica i gràfica entre dos o més mots que tenen significats diferents), entre altres.

Les classificacions que s'utilitzen en l'estadística oficial estableixen un conjunt jerarquitzat de categories homogènies que permeten classificar els elements que les componen. Aquests elements, en el nostre cas, poden ser les activitats econòmiques, les ocupacions i els béns i serveis (productes). Mitjançant aquests llenguatges estadístics (metainformació), les dades estadístiques es poden recollir, i una vegada tractades, es poden difondre d'una manera ordenada o classificada, amb més o menys nivells de desagregació o detall, per a uns àmbits determinats.

El TERMCAT ha col·laborat amb l'Idescat en l'elaboració de les principals famílies de classificacions estadístiques econòmiques (activitats econòmiques i productes) i de les classificacions demogràfiques i socials (ocupacions, educació i sanitàries) i l'elaboració de les quals ha generat un vocabulari molt extens en els diversos camps temàtics i alhora un enriquiment considerable de la terminologia catalana i la seva conseqüent aplicació a la realitat quotidiana, com ara l'etiquetatge de productes.⁴

Per tal d'exemplificar la feina feta, val a dir que el TERMCAT ha donat suport a l'elaboració de les principals classificacions d'activitats econòmiques (CCAE-93, CCAE-93 Rev.1 i CCAE-2009), classificacions de productes que deriven de les d'activitats econòmiques (CCPA-96, CCPA-96 Rev.1), classificacions de productes que són objecte de comerç amb l'estranger (Nomenclatura combinada), classificacions d'ocupacions (CCO-94), classificacions d'educació (CCED-2000), entre altres.

3 Excepcionalment, s'ha oficialitzat la *Classificació catalana de productes per activitats 1996 Revisió 1 (CCPA-96 Rev.1)*. Adaptació de la CPA-2002, derivada de la Classificació catalana d'activitats econòmiques 1993 Revisió 1 (CCAE-93 Rev.1), en motiu de la desagregació específica requerida per determinats agents del Sistema estadístic de Catalunya, malgrat no s'elaborés i normativitzés la versió espanyola de la classificació de productes, perquè l'INE va adoptar directament el Reglament comunitari CPA-2002.

4 Nomenclatures i classificacions estadístiques: <http://www.idescat.cat/cat/idescat/classif/>

Alguns exemples de consultes fetes al TERMCAT ha estat els següents:

Classificació catalana d'activitats econòmiques 2009 (CCAE-2009)

codi	forma consultada	proposta
20.30	llevavernissos	llevavernís

(Atès que el terme vernís és incomptable, la formació més adequada a l'hora de fer el compost és llevavernís i no llevavernissos. És un cas similar a formes com ara parabrisa, parafang, llevaneu, etc.).

Classificacions de les funcions de consum individual (COICOP_HBS)

codi	forma consultada	proposta
0511103	calzadora	cadira de calçar; cadira baixa

(La denominació castellana calzadora es documenta únicament en el *Diccionario del español actual* Manuel Seco per designar un "asiento bajo que se emplea para calzarse". Llevat d'aquesta font, no hem pogut documentar aquesta denominació en cap altra obra lexicogràfica o terminològica, en castellà o català. Com a alternativa catalana per fer referència a aquesta noció, es proposa el terme *cadira de calçar* o *cadira baixa* (forma recollida pel *Diccionari català-valencià-balear* per designar la cadira que té el seient més avall que les cadires ordinàries.)

Un altre exemple prou destacable de col·laboració el constitueix la Nomenclatura combinada, que inclou uns 20.500 epígrafs que despleguen unes 500.000 paraules i prop d'uns 100.000 termes, i, per tant, queda reflectit el volum de vocabulari treballat i la seva repercussió en l'àmbit del comerç amb l'estranger és gratificant.

La terminologia consultada i resolta es fa servir també a l'hora d'elaborar els qüestionaris de l'INE en format bilingüe, la qual cosa repercuteix en el fet que les empreses seleccionades en les mostres rebin els qüestionaris elaborats amb una terminologia precisa, senzilla, coherent i adequada a les seves funcions, i, que ben segur repercuteix en la qualitat de la resposta. Per tant, el bon ús de la terminologia incrementa sens dubte la qualitat de les operacions estadístiques.

4. Conclusions i reflexions generals

La celebració de les Jornades de Llengua i Estadística ens han permès als estadístics, lingüistes i terminòlegs, així com a tots els col·lectius que ens interessa l'evolució sòlida de la llengua, expressar i exemplificar que la interdependència que existeix entre la llengua i l'estadística és òbvia en l'àmbit de l'estadística oficial i, en aquest cas concret, esdevé una eina bàsica per elaborar els llenguatges estadístics.

Bàsicament, l'objectiu de l'article s'ha centrat a explicar la metodologia de treball pròpia que té l'Idescat a l'hora d'elaborar els llenguatges estadístics, remarcant que es compta amb un model de llengua propi de les nomenclatures i classificacions estadístiques que garanteix i contribueix en bona mesura a l'homogeneïtzació i la comparabilitat dels resultats estadístics, però alhora és un model de llengua que contribueix de forma creixent als processos de normalització de la llengua catalana.

I, amb aquest doble objectiu assolit en els nostres treballs quotidians, queda palès que l'estadística oficial catalana, a banda de donar compliment als objectius previstos a la Llei del Pla estadístic de Catalunya, reforça l'avanç de la llengua catalana per tal que sigui capaç de donar resposta als nous conceptes que es generen cada dia.

Val a dir que la definició d'un model de llengua propi és el resultat de la cooperació entre les diferents institucions i organismes que vetllem pel desenvolupament solvent de la nostra llengua, així com de l'important i imprescindible esforç que fan tots aquells agents del sistema estadístic als quals se'ls socialitza el llenguatge estadístic perquè donin la resposta més especialitzada possible en termes de precisió, atès que són els més competents per fer-ho. Resulta evident que la suma de tots els esforços ens permet obtenir un llenguatge estadístic de qualitat que, a més de ser un bon instrument estadístic, es caracteritza per la seva contribució continuada en l'avanç terminològic de la llengua del nostre país.

De fet, és ben cert que un país que progressa significa que la seva ciència, la seva tecnologia, la seva innovació, la seva recerca... avancen de forma ràpida, les quals han de poder disposar també d'una llengua que transmeti solvència i els permeti comunicar-se en una societat tan globalitzada com la que vivim. Al meu entendre, doncs, la llengua constitueix també un indicador del progrés d'un país, és a dir, no s'entendria que en un país que evolucioni econòmicament, socialment i políticament, no creixi lingüísticament i terminològicament parlant. El desenvolupament de la qualitat de la llengua és un clar indicador del desenvolupament d'una societat.

Després d'aquesta reflexió, tan sols voldria convidar-vos a tots els organismes implicats i als agents socials del nostre Sistema estadístic català a buscar plegats canals per tal de fer més visible i més explícita la llengua, en el sentit que esdevingui un indicador que mesuri, juntament amb molts d'altres (comparabilitat, confidencialitat, celeritat, homogeneïtat...) la qualitat d'una operació estadística, a banda, de mesurar intrínsecament, al cap i a la fi, l'evolució de la nostra societat.

Recursos terminològics per al treball estadístic

Glòria Fontova

TERMCAT, Centre de Terminologia

1. Introducció

Bon dia. En primer lloc, vull agrair als organitzadors que hagin tingut l'amabilitat de convidar el TERMCAT a participar en aquestes I Jornades de Llengua i Estadística. Abans d'entrar en matèria, i tenint en compte que la comunicació que m'ha precedit ha estat presentada per l'Institut d'Estadística de Catalunya, i que les tres comunicacions que m'han de seguir en aquest primer bloc aniran a càrrec de l'Instituto Nacional de Estadística, de l'Euskal Estatistika Erakundea, i de l'Instituto Galego de Estatística, crec que convé aclarir què hi fa un centre com el TERMCAT, que no correspon a cap organisme estadístic, enmig de l'Idescat, l'INE, l'Eustat o l'IGE.

Com a breu presentació als col·legues de fora de Catalunya i al públic assistent, voldria apuntar que el Centre de Terminologia TERMCAT, creat el 1985, és un organisme, integrat per la Generalitat de Catalunya, l'Institut d'Estudis Catalans i el Consorci per a la Normalització Lingüística, que té com a finalitat la promoció, la coordinació i el desenvolupament de les activitats terminològiques en llengua catalana. Des de fa més de dues dècades, el TERMCAT treballa per posar a disposició de la societat recursos terminològics de qualitat i per estandarditzar la terminologia científica i tècnica en català, amb el suport polític del govern de Catalunya, que garanteix l'extensió en l'ús de la terminologia desenvolupada, i amb el suport acadèmic de l'Institut d'Estudis Catalans que legitima les noves propostes terminològiques de normalització.

La nostra participació en aquest fòrum queda justificada, doncs, d'una banda, per l'interès terminològic que genera l'activitat estadística en el marc de la qual s'analitzen, es produeixen i es difonen dades sobre activitats econòmiques, productes, consum, salut, ocupacions i altres matèries que es vehiculen a través de la terminologia pròpia d'aquests àmbits. D'altra banda, la col·laboració existent des de fa molts anys entre l'Idescat i el TERMCAT, i prevista en el Pla estadístic de Catalunya, ens permet conèixer de ben a prop quins són els problemes terminològics que poden sorgir en els treballs d'actualització i desenvolupament de les nomenclatures i classificacions estadístiques oficials en llengua catalana.

Des d'aquestes dues coordenades, el desenvolupament i la difusió terminològiques, per una banda, i la col·laboració institucional, per l'altra, ens agradaria presentar en aquesta comunicació els recursos terminològics que el TERMCAT posa a disposició d'institucions i organismes que, com l'Idescat, treballen en la producció, adaptació i difusió de documentació especialitzada al mateix temps que es proposen garantir un ús adequat i uniforme de la terminologia utilitzada.

2. Recursos terminològics per a la comunicació especialitzada

En el marc de l'Administració pública, parlar de *recursos*, en plural, és fer referència al conjunt de mitjans humans, materials, informacionals i econòmics disponibles per al desenvolupament d'activitats planificades en els diferents àmbits d'actuació competencial. En l'àmbit de la terminologia, l'Administració pública ha anat generant múltiples recursos per impulsar i garantir l'ús adequat de la llengua i la terminologia en les comunicacions produïdes tant en els òrgans administratius com en els diferents segments de la vida social. La xarxa d'assessors lingüístics adscrits als diferents departaments del Govern o a organismes com l'Idescat és un exemple representatiu de recursos terminològics creats amb l'objectiu de prestar l'assessorament lingüístic necessari i garantir l'ús correcte i uniforme de la terminologia en el si dels seus respectius organismes.

En aquesta mateixa línia, el TERMCAT esdevé un recurs per als usuaris que s'hi adrecen per a resoldre un dubte terminològic, per a rebre assessorament en l'elaboració d'un diccionari, d'una norma tècnica, d'una classificació estadística o d'un programa informàtic, per a obtenir informació documental sobre un determinat àmbit del coneixement o per a sol·licitar l'estudi d'un neologisme o la revisió d'una forma normalitzada. Per a atendre a les diferents necessitats expressades per les institucions i organismes, tant dels sectors públics com privats, i per la societat en general, el TERMCAT ofereix un conjunt de recursos terminològics entre els quals destaquen:

1. Diccionaris i altres reculls terminològics que apleguen la terminologia dels diferents sectors d'especialitat.
2. Assessorament a institucions, organismes i persones per a l'elaboració de diccionaris terminològics, per a la redacció o revisió de documentació especialitzada i per a la localització de productes informàtics.
3. Servei d'atenció i resolució de consultes terminològiques o documentals de qualsevol àmbit d'especialitat.
4. Normalització dels neologismes del lèxic especialitzat en llengua catalana.
5. Criteris i mètodes aplicats al treball terminològic.
6. Aplicacions i eines gratuïtes per a:
 - la consulta terminològica multilingüe
 - la gestió del treball terminològic

3. Recursos terminològics per al tractament de les classificacions estadístiques

El procés de traducció i adaptació de les classificacions estadístiques d'una llengua a una altra sol ser complex, atesa la multidisciplinarietat que les caracteritza, el lèxic especialitzat que s'hi utilitza i el volum de vocabulari que contenen. La qualitat lingüística que ha de prevaler en aquest procés ha d'estar especialment garantida pel fet que les classificacions estadístiques, un cop finalitzat el procés d'oficialització, es difonen en múltiples sectors i són d'ús obligatori en l'àmbit de l'Administració pública.

En les actuacions que l'Idescat du a terme per adaptar i oficialitzar les nomenclatures i classificacions estadístiques, el procés de traducció al català hi té un paper rellevant. Els assessors lingüístics de l'Idescat s'encarreguen de revisar els treballs de traducció de les classificacions i d'establir

els criteris que defineixen el model de llengua utilitzat en aquests documents. Generalment, és en la fase de revisió de la versió catalana d'una classificació o nomenclatura que solen sorgir dubtes terminològics que requereixen la participació del TERMCAT. D'altres vegades, l'Idescat sol·licita al Centre de Terminologia la revisió terminològica final d'una proposta de classificació abans d'iniciar-ne el procés d'oficialització.

Tant en un cas com en l'altre, els principals recursos terminològics destinats al tractament d'aquests instruments estadístics són l'assessorament terminològic i la normalització dels neologismes que ho requereixen. L'assessorament té com a objectiu oferir una proposta denominativa adequada per a la resolució d'un dubte terminològic, mentre que la normalització entra en joc quan es considera convenient fixar una forma lingüística per a denominar un concepte determinat.

3.1 Assessorament terminològic

Els treballs d'assessorament terminològic per a l'elaboració d'una classificació comporten, en primer lloc, el tractament d'un gran volum de terminologia. El cas de la *Nomenclatura combinada*, amb 20.500 epígrafs agrupats en 100 capítols que despleguen unes 500.000 paraules i prop de 100.000 termes, pot donar una idea de la complexitat que genera la gestió terminològica d'un producte d'aquestes característiques. En el treball de revisió d'aquesta nomenclatura, el TERMCAT va assessorar l'Idescat en la resolució de més de 400 dubtes terminològics.

En segon lloc, el treball terminològic en aquest terreny va lligat indefectiblement a la recerca multidisciplinària. La *Nomenclatura combinada*, exemple rellevant també en aquest cas, va generar dubtes terminològics relacionats amb espècies pesqueres, fustes tropicals, formatges, teixits, aparells de mesura, productes químics i molts altres àmbits d'especialitat. El coneixement permanentment actualitzat dels recursos terminològics multidisciplinaris existents es converteix, doncs, en un requisit bàsic en els treballs d'assessorament orientats a aquests productes.

Entre les principals fonts de consulta que cal tenir en compte en la revisió de la versió catalana d'una classificació convé remarcar, com a punt de partida, l'accés a les versions de la classificació en altres llengües, especialment el castellà, l'anglès i el francès, però també, en els casos que ho requereixen, l'italià, el portuguès i l'alemany. La comparació entre les diferents versions lingüístiques d'una classificació permet delimitar correctament el terme objecte d'estudi i conèixer-ne els equivalents en altres llengües, detectar possibles errors de traducció, identificar la naturalesa lèxica del terme dubtós, adscriure'l a un àmbit temàtic concret i situar-lo en un context d'ús.

Un altre bloc de recursos terminològics el constitueixen les obres lexicogràfiques considerades de referència bàsica en les diferents llengües de treball, la majoria de les quals actualment són consultables en línia de manera gratuïta. El *Diccionari de la llengua catalana*, de l'Institut d'Estudis Catalans, és un dels corpus de consulta obligada que ajuda a resoldre un bon nombre dels dubtes terminològics que sorgeixen en la revisió de classificacions i nomenclatures estadístiques. A les fonts lexicogràfiques, cal afegir-hi, també, el recurs a les fonts terminològiques i especialitzades més representatives dels diferents àmbits del coneixement a què pertanyen els termes objecte de consulta.

Les bases de dades, els diccionaris electrònics i, en general, els recursos multilingües en línia que ofereixen la possibilitat de consultar la terminologia pròpia dels diferents àmbits d'especialitat són també elements imprescindibles en la recerca terminològica. El Cercaterm, el servei de consultes terminològiques en línia del TERMCAT, permet la consulta de tots els termes normalitzats pel Consell Supervisor, de la terminologia procedent dels productes terminològics que han estat elaborats o assessorats pel Centre, dels projectes de recerca sectorial que estan en un estat avançat d'elaboració i de les recerques puntuals dutes a terme per les diferents àrees de treball.

Finalment, també cal considerar com a recurs terminològic la informació terminològica localitzable a Internet. Utilitzar la xarxa com a font de consulta permet extreure, de manera gairebé instantània, informació sobre un terme. Ara bé, en aquest cas, més que no pas en cap altre, caldrà extreure la ponderació sobre la fiabilitat i la representativitat dels documents consultats.

Quan les fonts de consulta que s'han esmentat fins ara no són suficients per a resoldre satisfactòriament la qüestió que ha motivat la recerca, el recurs als especialistes de l'àmbit temàtic a què pertany el terme objecte d'estudi es fa totalment imprescindible. En el procés de revisió terminològica de la versió catalana de la *Nomenclatura combinada*, es va haver de recórrer a especialistes dels àmbits de la botànica, la zoologia, l'enologia, la química, la indústria del tabac, del paper, de la confecció, del vidre o de la rellotgeria, entre molts altres, per a la resolució d'una bona part dels dubtes terminològics derivats de la traducció. La seva intervenció va ser útil a l'hora d'aclarir conceptes, d'elaborar una proposta denominativa, de valorar la viabilitat d'un neologisme o, fins i tot, de fer d'intermediaris amb altres especialistes del seu sector.

3.2 Normalització terminològica

Pot passar que, després d'exhaurir tots els recursos que s'utilitzen en el treball d'assessorament terminològic, la recerca s'acabi sense poder resoldre el dubte inicial. En alguns casos, es poden detectar termes no documentats en català i designats amb formes manllevades (*slack wax*, *top-cross*, *crumb*) o amb calcs d'altres llengües (*dead burning*: *calcinació a mort*). En d'altres casos, el problema sorgeix quan un mateix terme es difon amb denominacions diferents (*foca de cresta* i *foca de casc*), o bé quan cal establir criteris generals per a grups de termes que comparteixen una determinada característica (com és el cas, per exemple, de la denominació dels formatges en català). També es pot donar el cas en què calgui reconsiderar una proposta normalitzada prèviament, perquè hi ha arguments prou justificats, sovint aportats pels mateixos especialistes o usuaris, que recomanen replantejar una determinada solució.

En el cas concret de la *Nomenclatura combinada*, dels 400 dubtes terminològics que va generar la revisió de la versió catalana d'aquesta obra, gairebé en un centenar de casos es va haver de recórrer al recurs de la normalització per poder fixar una alternativa denominativa amb prou garanties d'adequació i de consens. Per a cadascun dels termes estudiats, es va elaborar un dossier de normalització, és a dir un document que, de manera metòdica i sistemàtica, recollia totes les informacions sobre la recerca terminològica duta a terme i sobre els aspectes lingüístics, terminològics i sociolingüístics que afectaven cada cas.

El dossier de normalització és la base sobre la qual el Consell Supervisor pot valorar la informació recollida i prendre la decisió adequada. Aquest òrgan, format per membres de la Secció Filològica de l'Institut d'Estudis Catalans, per membres del TERMCAT i per especialistes designats per aquests dos organismes, s'encarrega de la fixació dels neologismes terminològics en llengua catalana i es reuneix quinzenalment amb la finalitat d'adoptar resolucions sobre els casos estudiats o sobre criteris aplicables al treball terminològic.

Entre les actuacions de normalització més significatives dutes a terme pel Consell Supervisor en el projecte de la *Nomenclatura combinada* destaquen l'aprovació de 16 denominacions d'espècies pesqueres, de 36 denominacions de formatges o de 40 denominacions de fustes tropicals. En aquests dos darrers casos, el Consell Supervisor va definir, a més a més, criteris específics que van permetre el tractament homogeni de les denominacions estudiades.

4. Difusió de la terminologia i actualització dels recursos terminològics

Al llarg d'aquesta exposició s'ha volgut fer èmfasi en dos dels recursos terminològics de què disposa l'Idescat a l'hora de treballar en el procés d'adaptació i oficialització de les classificacions i nomenclatures en llengua catalana: l'assessorament i la normalització. En tots dos casos, s'ha vist com de la utilització d'aquests recursos es deriven solucions terminològiques que retornen al cos de la classificació o de la nomenclatura. Així, doncs, l'element que inicialment ha suscitat la demanda i el desplegament de recursos terminològics, es converteix al final del procés en una de les principals vies de difusió de la terminologia tractada. Paral·lelament, aquesta terminologia també s'incorpora a la Neoloteca, en el cas dels termes normalitzats pel Consell Supervisor, al Cercaterm i als altres recursos consultables o descarregables lliurement des del lloc web del TERMCAT. Al mateix temps, la terminologia generada passa a formar part del corpus de treball per a l'elaboració de nous diccionaris i reculls terminològics.

La nostra experiència de cooperació amb els sectors de l'Administració, les universitats, els mitjans de comunicació, les empreses, els col·legis professionals o les associacions, ens permet confirmar que la sinergia entre els diversos agents (especialistes, lingüistes, redactors, etc.) que incideixen en els models lingüístics d'una comunitat, afavoreix la qualitat dels llenguatges especialitzats i l'actualització constant dels recursos terminològics disponibles.

Per acabar, i fent-me disculpar el meu nul coneixement en matèria estadística, m'agradaria posar en una gràfica els elements que acabem de comentar: recursos terminològics, partenariats i bones pràctiques en el treball estadístic i terminològic. De la combinació d'aquestes tres variables es podria obtenir l'evolució de la salut lingüística en l'activitat estadística, objecte de debat en aquest primer bloc de la Jornada. A mi se'm fa difícil imaginar un altre resultat que no sigui un increment sostingut de la qualitat de vida de la llengua en l'àmbit de les estadístiques oficials.

Moltes gràcies.

El lenguaje estadístico y el inglés

Francisco Hernández Jiménez⁵

Subdirector General de Metodología y Técnicas Estadísticas

En esta presentación se reflexiona sobre la importancia del inglés en la estadística oficial. Si bien a nivel científico la estadística apoya su formulación en el lenguaje matemático, en la estadística oficial además es muy importante la comparabilidad y armonización de la información, y para ello se debe garantizar la coherencia de las definiciones de los diferentes conceptos a considerar. El lenguaje de referencia internacional es el inglés, y en esta presentación se va a exponer cómo el mismo se ha convertido en una herramienta fundamental en la estadística.

1. El lenguaje científico

La estadística como ciencia se soporta sobre el lenguaje matemático. Esto ha permitido que se hayan difundido los avances científicos en esta materia con una enorme rapidez y precisión. Así, la formulación matemática de la estadística ha facilitado que los conceptos de la misma hayan podido ser claros y precisos para la comunidad científica. Faltaba por asegurar que la denominación de estos conceptos en todos los idiomas tuviera garantizada su correcta interpretación. Para ello, el ISI (*International Statistical Institute*), a través del IAOS (*International Association for Official Statistics*), ha documentado un vocabulario de términos comunes en estadística en diferentes lenguas, que garantiza la interpretación coordinada en todos los idiomas.

Las formas principales de comunicación en la comunidad científica actualmente son las publicaciones y revistas especializadas, los congresos y seminarios que se organizan, y sobre todo la difusión en Internet, que abre un amplio espectro de posibilidades en términos de rapidez y accesibilidad.

El idioma de mayor difusión en todos estos medios es el inglés por lo que, en la medida de lo posible, se debe intentar adquirir conocimientos y ofrecer experiencias e investigaciones a otros expertos utilizando este idioma. No obstante, los resultados de nuestras investigaciones se deben compartir en otras lenguas, intentando abarcar el máximo de la comunidad científica.

2. Influencia del inglés en la comparabilidad y la armonización estadística

Una de las características básicas de los datos de estadística oficial es que han de responder a las necesidades de la sociedad, y una de las fundamentales en un mundo cada vez más globalizado es la comparabilidad de la información. Como ejemplo de este hecho se puede señalar el *informe PISA*, cuyo objetivo es comparar entre los países de la OCDE cuál es el nivel de conocimiento de los alumnos de secundaria.

⁵ Este artículo refleja opiniones personales del autor. En ningún caso es un reflejo de la opinión oficial del Instituto Nacional de Estadística.

Los trabajos realizados en la estadística oficial requieren, por un lado, del intercambio de conocimientos y experiencias entre los expertos de los diferentes organismos y países, y por otro, de la transmisión de la información a los usuarios de forma clara y concisa. Cuando el intercambio o la difusión se hace a nivel nacional el idioma utilizado es el propio de ese país o nación, pero cuando se traspasan las fronteras nacionales el idioma universal es el inglés.

El uso del inglés con carácter generalizado en la comunicación entre los expertos y entre las oficinas y los usuarios de los distintos países, hace necesario ser muy cuidadosos con la terminología que se utiliza, teniendo en cuenta los principios básicos de armonización y precisión. De un mal uso del idioma o de una mala traducción se pueden derivar problemas en la toma de acuerdos o en la toma de decisiones que pueden padecerse durante varios años, o bien, como veremos a continuación, pueden quedar en anécdotas curiosas.

2.1. Textos legales

La Comisión Europea tiene la obligación de que todos los Reglamentos, antes de ser publicados en el Diario Oficial de la Unión Europea, sean traducidos por su departamento de traducción a las lenguas oficiales reconocidas por la UE. Dicho departamento, en contactos bilaterales con los diferentes Estados miembros, acuerda una traducción de los textos jurídicos. Dado que se lleva trabajando muchos años de esta forma, no existen grandes problemas en la traducción del articulado o la exposición de motivos cuando se refiere a términos jurídicos. Los problemas pueden aparecer cuando ese texto jurídico se refiere a las definiciones a utilizar, normalmente referidas a lenguajes técnicos sobre los que los juristas lingüistas tienen una menor experiencia.

A modo de ejemplo, veamos un poco más extensamente los problemas que pueden aparecer al realizar la traducción de la estructura de una clasificación, cuyos títulos deben estar caracterizados por la precisión y claridad.

En la última revisión de las clasificaciones económicas europeas, denominada Operación 2007, la Comisión Europea no admitió a los Estados miembros traducciones de los títulos de la NACE Rev.2⁶ que no fuesen estrictamente literales⁷ del inglés. El objetivo era la armonización, es decir, que un mismo código tuviese el mismo significado en todos los Estados miembros. Esta exigencia retrasó mucho los trabajos de traducción, porque en el caso de los países que comparten lengua, cada uno hizo una traducción adaptada a las peculiaridades de su país, que fueron rechazadas al enfrentarse a versiones diferentes para un mismo idioma (por ejemplo, alemán o francés). La Comisión obligó a que la traducción al francés y al alemán fuese idéntica en todos los Estados miembros que hablan esos idiomas. Aunque en España no tuvimos ese problema, la búsqueda de armonización por parte de la Comisión llevó a la pérdida de precisión en los títulos en español, y en algunos casos, a absurdos que hubo que poner de manifiesto y tratar de buscar una solución.

Al verificar la traducción de la Comisión en la unidad de clasificaciones del INE se tuvo en cuenta como origen de la traducción el texto en inglés, la traducción de la versión anterior de la clasificación, la NACE Rev.1.1, y la opinión de los expertos en los diferentes sectores económicos y de los diferentes usuarios de la clasificación, como las CCAA, siendo la mayoría de estos expertos conocedores de la clasificación, pues formaron parte del grupo que España había creado para su elaboración. Entre los sectores económicos más sensibles al realizar la traducción del inglés al español cabe destacar el sector financiero, la educación o la asistencia social entre otros.

⁶ Nomenclatura de Actividades Económicas de las Comunidades Europeas, Revisión 2.

⁷ En la última revisión del 1993 hubo bastante discrepancia entre las versiones nacionales y la versión europea en inglés, y desde entonces tanto las versiones europeas como las posteriores adaptaciones nacionales, que han de ser aprobadas por la UE, están sometidas a un estricto control.

A continuació figuren alguns exemples ilustratius de los problemes encontrats:

- En alguns casos el problema consistí en un uso inadequat del idioma, possiblement per la utilització de un traductor electrònic en una primera versió. Per exemple, per el còdig *55.3 Camping grounds, recreational vehicle parks and trailer parks*, la Comissió proposa com a traducció *Càmpines*, paraula que no existe en espanyol. El acord final fue traduirlo com *Campings y aparcamientos para caravanas*.
- Para una mejor comprensió del contenido de la categoría *01.3 Plant propagation*, el grupo de clasificaciones del INE proposa traduirlo com *Viveros y cultivos de plantas ornamentales*. Dicha propuesta no fue aceptada por la Comissió y, después de numerosas consultas con los diferentes expertos, se aceptó *Propagación de plantas*, aunque el título en espanyol no sea el mejor.
- La imposibilidad de realizar la traducción de determinados términos, por ser muy novedosos, o el uso de anglicismos en ciertos sectores específicos lleva a que éstos sean aceptados en la clasificación. Un ejemplo es la aceptación del término *hosting* en el sector de las tecnologías de la información, o *holding* en el financiero.
- Se intentó evitar la pérdida de precisión en la traducción de los títulos. Por ejemplo, la categoría *62.0 Computer programming, consultancy and related activities* fue traducida por la Comissió com *Servicios de tecnologías de la información*, título demasiado amplio que no ayudaba a delimitar el contenido de la categoría. Finalmente se acordó *Programación, consultoría y otras actividades relacionadas con la informática*.
- Una traducción errónea o imprecisa de un término en inglés puede generar confusión entre distintas categorías de la clasificación. Por ejemplo, las propuestas de traducción de la Comissió para las categorías *64.91 Financial leasing* y *77 Rental and leasing activities* eran *64.91 Arrendamiento financiero* y *77 Actividades de alquiler y arrendamiento financiero* respectivamente, generando un problema con la clasificación del arrendamiento financiero. La Comissió siempre traducía la palabra *leasing* com arrendamiento financiero, ya que el sector utiliza esta palabra para designar dicho producto. Sin embargo, en inglés el arrendamiento financiero se denomina *financial leasing*, debiéndose traducir el término *leasing* com alquiler o arrendamiento.
- En el sector de los servicios sociales se tuvo una especial sensibilidad con el lenguaje que se utilizaba, adaptándose al utilizado en el *Libro Blanco de la dependencia*. Así por ejemplo, se evitó traducir los términos *elderly* y *mental retardation* com ancianos o retraso mental, utilizándose los términos de *personas mayores* y *discapacidad intelectual* respectivamente.

2.2. Documentos metodológicos

Los documentos metodológicos son documentos de apoyo, originalmente en inglés, cuya redacción se acuerda entre los Estados miembros. La traducción de estos documentos es responsabilidad de dichos Estados, permitiéndose una mayor flexibilidad o adaptación del lenguaje, sin olvidar que su objetivo es garantizar la armonización estadística.

Siguiendo con el ejemplo de las clasificaciones, podemos mencionar com documentos metodológicos asociados a una clasificación sus notas explicativas y su introducción metodológica.

Las notas explicativas son una herramienta esencial para los usuarios, que permite el uso de la clasificación de forma armonizada y precisa. Explican en detalle el contenido de cada una de las categorías de la clasificación y, por lo tanto, incluyen descripciones muy específicas de actividades, productos u ocupaciones, lo cual dificulta mucho su traducción.

Una buena traducción se consigue con el uso de la clasificación a partir de una primera traducción provisional de las notas, y con las mejoras aportadas por los usuarios y los expertos en el sector. En el sector financiero es muy importante la cooperación con el Banco de España, el cual ha considerado que el uso de anglicismos para ciertos productos financieros es más adecuado, por ejemplo, *swaps*. En algunos casos se ha considerado que es conveniente incluir entre paréntesis la denominación en inglés para aclarar la traducción al español, por ejemplo, “fondos de inversión colectiva (*unit investment trust funds*)”. Esta misma solución se ha tomado en otros sectores, como el de las nuevas tecnologías: “puentes (*bridges*), enrutadores (*routers*) y puntos de acceso (*gateways*)”.

Algunos anglicismos utilizados están reconocidos por la Real Academia de la Lengua, pero otros no. Entre los primeros podemos encontrar ejemplos como *marketing*, *software*, *offset* (impresión *offset*), *input* o *web*. Algunos ejemplos de anglicismos no reconocidos son *fitness*, *intranet*, *piercing* o *joystick*. Incluso hay palabras que sin ser de uso habitual en inglés, son utilizadas por los especialistas de un determinado sector, y en aras de facilitar el uso de la clasificación en estos casos, se incluyen en las notas explicativas, como es el caso de *vending* o *footing*.

En numerosas ocasiones se incluyen también siglas que proceden del inglés y que son de uso generalizado y no se traducen, como *SMS* (*Short Message Service*), *CD* (*Compact Disc*), *DVD* (*Digital Video Disc*) o *GPS* (*Global Positioning System*). Sin embargo, no existe una norma general para la traducción de las siglas. En algunas ocasiones se traducen por las siglas equivalentes en español, como *ICT* (*Information and Communication Technologies*) que se traduce como *TIC* (*Tecnologías de la Información y las Comunicaciones*), o por una descripción en español, como *ATM* (*Automatic Teller Machine*), traducido como *cajero automático*.

La introducción metodológica de las clasificaciones define los principios básicos para la correcta clasificación de las unidades estadísticas y, por lo tanto, no puede resultar ambigua en su interpretación. Por ejemplo, en la redacción de la nota introductoria de la NACE Rev.2 surgió un problema con la interpretación de ciertos términos relacionados con la subcontratación (*outsourcing*). Supongamos que una unidad A subcontrata parte de su proceso de producción con otra unidad B ¿Cómo denominamos a las unidades A y B?. Algunos expertos usaban el término *principal* para referirse a la unidad A, y *contractor* para la unidad B. Otros denominaban *contractor* a la unidad A y *subcontractor* a la unidad B. Para resolver este problema se acordó denominar *principal* a la unidad A y *contractor* a la unidad B, incluyendo una definición de cada uno de estos términos en la introducción metodológica.

2.3. Documentos de trabajo

Existen ciertos ámbitos estadísticos emergentes, cuyos desarrollos científicos no están basados en el lenguaje matemático, para los que aún no existe una normalización, ni siquiera en inglés. Como veremos a continuación, esto condiciona mucho la comunicación entre países. En estos casos hay que ser mucho más cuidadoso en el uso del lenguaje, tanto en inglés como en español.

Un ejemplo de este tipo lo podemos encontrar en el ámbito de los metadatos estadísticos. Esta parte de la estadística oficial es relativamente moderna. Desde hace unos 9 años existen grupos de trabajo europeos sobre metadatos en diferentes ámbitos, como la informática, la difusión o la metodología. El hecho de que estos grupos hayan trabajado de forma independiente ha llevado a que el lenguaje utilizado difiera de un grupo a otro. Por ejemplo, en los grupos de trabajo de difusión se utiliza la denominación de *división* para lo que en metodología llamamos *statistical operation*. Esto dificulta la visión global de los metadatos y el intercambio de experiencias entre grupos. Para afrontar este problema se está desarrollando el denominado *Common Metadata Vocabulary*, que es un glosario de términos relacionados con los metadatos.

Este vocabulario común no sólo permitirá que un mismo concepto se designe con un único término, sino que también evitará que un término se interprete con distintos significados. Por ejemplo, el término *classification* se utiliza en algunos casos sólo para las clasificaciones aprobadas por un Reglamento europeo (por ejemplo, la NACE Rev.2), y en otros casos se utiliza para designar cualquier lista de los posibles valores de una variable (por ejemplo, estado civil). Este problema salió a la luz cuando se cuantificó el número de *classifications* consideradas por los países en sus sistemas de metadatos, en el primer caso no más de 10 y en el segundo alrededor de 400.

En otras ocasiones los problemas están relacionados con los diferentes marcos legales y culturales de los países. Así, las distintas figuras jurídicas que pueden presentar las empresas (persona física, sociedad anónima, sociedad de responsabilidad limitada, sociedad cooperativa, asociación, fundación, etc.) pueden diferir mucho de las existentes en otros países, haciendo muy difícil la comparación internacional de por ejemplo, la demografía empresarial desagregada por forma legal. Pretender unificar realidades jurídicas diferentes en un mismo término en inglés es un *atrevimiento* en el que a veces se incurre al querer llevar la comparabilidad y la armonización más allá de lo posible.

3. Desarrollo de las reuniones internacionales

Desde el punto de vista del idioma existen dos tipos de reuniones internacionales: las que cuentan con traducción simultánea y las que se desarrollan en inglés. Aunque en un principio pueda parecer que las primeras son más sencillas, a continuación se presentan algunos problemas sufridos en ambos tipos de reuniones.

En las reuniones con traducción nos enfrentamos a dos problemas distintos. Por un lado, cuando los participantes pueden hablar en su lengua materna tienden a hacer exposiciones más largas, y por lo tanto más difíciles de traducir y comprender. En segundo lugar, y hablando específicamente de las reuniones sobre temas estadísticos, hay que tener en cuenta que los traductores no suelen ser expertos en el tema. En el caso de Eurostat son traductores del departamento lingüístico del Parlamento, acostumbrados al lenguaje parlamentario y no al técnico. Esto provoca que las traducciones no tengan la suficiente calidad para garantizar un buen entendimiento, perdiéndose mucha precisión en la comunicación. La ventaja obvia es que aquellos expertos en los temas tratados que carecen de conocimientos en inglés puedan participar en las reuniones, manifestando sus opiniones.

En cuanto a las reuniones sin traducción, los participantes intentan ser más concisos en las ideas y cada experto utiliza el inglés a su nivel, evitándose en lo posible por aquellos con un nivel más alto, como los nativos, los giros lingüísticos que obstaculizan la comunicación. En estas reuniones se hace uso del ingenio para que se entiendan las ideas principales, y una mala pronunciación o la falta de vocabulario se solucionan con imaginación, por ejemplo a través de dibujos o símiles.

4. Siglas propias del lenguaje estadístico

En el apartado 2.2 se han presentado algunos ejemplos del uso de siglas en diferentes documentos estadísticos, que se corresponden con siglas utilizadas en el lenguaje habitual de la población.

Además existen otras siglas que son propias del lenguaje estadístico, conocidas y utilizadas a veces de forma abusiva por la comunidad científica, como *LFS* (*Labor Force Survey*, que en español se corresponde con la EPA), *GNP* (*Gross National Product*, PNB en español), o en general las siglas asociadas a cualquier encuesta o proyecto estadístico. Algunas son de uso común para la población en general, como el IPC.

Incluso hay otras siglas que se refieren a nombres de clasificaciones, una de cuyas principales funciones es acompañar a los datos en su difusión, optándose a nivel europeo por utilizar en todos los idiomas las mismas siglas, aunque no se correspondan exactamente con las iniciales de esa clasificación en el idioma utilizado.

Así, la NACE⁸ se corresponde con las siglas en francés de *Nomenclature statistique des Activités économiques dans la Communauté Européenne*, que es la clasificación de actividades económicas europea, y esta sigla se usa en todos los idiomas, habiéndose traducido el título de la clasificación.

Precisamente el abuso, y a veces mal uso de las siglas de clasificaciones, puede convertir a éstas en un problema. Toda sigla, ya fuera a pie de página o de otra forma, debería ser explicada en el propio texto. Si la sigla está en inglés, tal vez fuera conveniente poner su texto en inglés y su traducción al idioma utilizado en ese momento. El mal uso de las siglas se produce cuando no se utiliza la sigla en toda su extensión, por ejemplo, en el caso de las clasificaciones cuando se pierde la referencia temporal. El INE recibe algunas consultas sobre siglas, y por lo general se trata de siglas que acompañan a datos, observándose que en muchas ocasiones las siglas son incompletas, especialmente faltando la referencia temporal.

Las siglas también están sometidas al problema de la traducción. Por citar un problema de una traducción no precisa, a nivel de Naciones Unidas las siglas de algunas clasificaciones se han traducido a las diferentes lenguas, y no siempre acertadamente. Por ejemplo, la sigla *ISIC* se corresponde con la *International Standard Industrial Classification of all economic activities*. Cuando esta sigla se tradujo al español se optó por la sigla *CIIU*, que corresponde a *Clasificación Industrial Internacional Uniforme de todas las actividades económicas*. Al no formar parte de las siglas, la expresión “de todas las actividades económicas”, no se suele incluir. El usuario puede interpretar que se trata de una clasificación de la industria manufacturera, cuando el término inglés *industry* se debería traducir como sector económico.

5. Conclusiones

- La estadística a nivel científico se apoya en el lenguaje matemático, y por tanto, no tiene especiales problemas para una interpretación correcta.
- En la estadística oficial, además hay que asegurar la comparabilidad de la información, tanto a nivel nacional como internacional, lo que obliga al uso de conceptos y definiciones comunes, y de una terminología común en el lenguaje internacional, el inglés.
- El uso del inglés como instrumento de trabajo, dependiendo del ámbito, presenta problemas diversos. Es muy importante alcanzar acuerdos para el uso de vocabularios comunes que faciliten las tareas de comparabilidad.
- Por último, la buena voluntad por parte de todos los implicados en el proceso de elaboración de la estadística oficial, facilita a nivel internacional la comprensión y avance en todos los proyectos, y como ejemplos se pueden destacar los logros alcanzados en la Unión Europea en esta materia, lo que ha permitido crear un sistema estadístico europeo fiable y comparable.

8 Esta clasificación data del 70 y entonces el francés era la lengua de uso más frecuente en la UE.

Referencias

- **Common Metadata Framework:** United Nations Economic Comision for Europe: <http://www.unece.org/stats/cmf/>
- **Common Metadata Vocabulary:** SDMX (Banco Central Europeo, Eurostat, Fondo Monetario Internacional, OCDE, Naciones Unidas, Banco Mundial): http://sdmx.org/wp_content/uploads/2008/02/sdmx_annex4_metadata_common_vocabulary_draft_february_2008.doc
- CPA 2008:
 - Reglamento (CE) No 451/2008 del Parlamento Europeo y del Consejo de 23 de abril de 2008 que establece una nueva clasificación estadística de productos por actividad (CPA) y modifica el Reglamento del Consejo (EEC) No 3696/93
 - Notas explicativas: <http://circa.europa.eu/irc/dsis/nacecpacon/info/data/en/CPA%202008%20structure%20and%20explanatory%20notes%20-%20EN.pdf>
- NACE Rev.2:
 - Reglamento (CE) No 1893/2006 del Parlamento Europeo y del Consejo de 20 de diciembre de 2006 por el que se establece la nomenclatura estadística de actividades económicas NACE Revisión 2 y por el que se modifica el Reglamento (CEE) no 3037/90 del Consejo y determinados Reglamentos de la CE sobre aspectos estadísticos específicos
 - Notas explicativas: <http://circa.europa.eu/irc/dsis/nacecpacon/info/data/en/NACE%20Rev%202%20structure%20+%20explanatory%20notes%20-%20EN.pdf>
 - Documento metodológico: <http://circa.europa.eu/irc/dsis/nacecpacon/info/data/en/NACE%20Rev.%202%20Introductory%20guidelines%20-%20EN.pdf>

Glosario del web de Eustat

El miedo al botijo produce monstruos⁹

Joxean Sagastizabal
Traductor de l'Eustat

Si tuviera que ponerle un segundo título a este trabajo, no lo dudaría, sería: **EL MIEDO AL BOTIJO PRODUCE MONSTRUOS.**

Probablemente lo parece, pero no es una “boutade”; evidentemente, requiere una explicación.

Se atribuye a Ibon Sarasola, muy reputado lingüista vasco (que, por cierto, vivió y trabajó muchos años en Barcelona) el uso de la palabra “botijo” para referirse a los calcos lingüísticos de palabras que sólo se dicen en castellano. Alguien estaría incurriendo en “botijismo” si, al buscar una palabra adecuada para traducir, por ejemplo, “quirófano” al euskera, pusiese “kirofanoa”, sin preocuparse de mirar un poco en las lenguas de alrededor. Si lo hiciera, vería que “quirófano” es “operating room” o “operating theatre” en inglés, “salle d’operation” en francés y “operationssaal” en alemán. Entonces, al traductor se le podría ocurrir, como efectivamente sucedió en su momento, que en euskera se podría utilizar “ebakuntza-gela”.

Supongo que lo que acabo de exponer suena bastante lógico, pero el “antibotijismo” digamos, radical, que fue mayoritario en el País Vasco durante los años ochenta y primeros noventa, produjo monstruos, textos absolutamente farragosos, difícilísimos de entender, escritos bajo la orden de que “nada debe decirse como en castellano”. Un buen ejemplo es la Ley 4/1986, de 23 de abril, de Estadística de la Comunidad Autónoma de Euskadi; un fenómeno, lingüísticamente hablando, en el que uno encuentra palabras como: egiterapena, izaiki, herrilanari, zertazkoa, arioz, irozkaia, eraentza, iman-irarki y un larguísimo etcétera, palabras cuya característica fundamental es que suenan a chino a las personas que desconocen el euskera y también a los que lo saben perfectamente. Y es esta ley la que en su artículo 28 crea el Euskal Estatistika Erakundea/Instituto Vasco de Estadística, el Eustat, criatura hija de una ley perfectamente incomprensible.

Una anécdota personal de los años “radicales antibotijistas” puede ilustrar aún más el desbarajuste léxico-sintáctico de la época. Yo entonces era profesor de euskera en una academia pública para adultos, en un euskaltegi, junto con unos veinte colegas. No sé qué día de no sé qué mes de no sé qué año, a principios de los noventa, recibimos todos un escrito oficial, que, siendo sus receptores la flor y nata de la cultura vasca, venía sólo en euskera (cosa en absoluto habitual, ni entonces ni hoy en día). Pues bien, alguien, algún héroe sincero, se atrevió a decirlo: “No entiendo nada, ¿qué es esto? ¿Qué se nos comunica oficialmente? Y, como todos estábamos igual, nos reunimos y después de muchos esfuerzos para descifrar aquel texto radical antibotijista entendimos que aquello era nada menos que un aviso de apertura de expediente disciplinario (por un día de huelga *salvaje*).

9 *Nota d'Edició:* El títol primer que d'entrada va acceptar el ponent va ser *Glosario del web de Eustat*. Atès, però, el fil argumentatiu de la ponència, creiem adient fer-hi constar el segon títol, que és el que realment justifica la seva intervenció.

Nadie debe pensar que la ideología superpurista de la época exigiera que aquel euskera marciano fuera lo que les enseñábamos a nuestros alumnos. No, el antibotijismo se centraba en los textos, fundamentalmente en los oficiales traducidos del castellano.

Se supone que no hay mal que cien años dure, y las cosas empezaron a cambiar. Hay una publicación que marca un antes y un después, un claro hito. Me refiero a Hizkera Argiaren Bidetik, que traducido es algo así como “Por un lenguaje claro”, publicado por el Instituto Vasco de Administración Pública (IVAP) en 1994. Es un libro lleno de sentido común, cuyo prólogo dice, en su última frase: “... a ver si entre todos somos capaces de emplear en la administración un euskera correcto y comprensible.” Una gran idea.

No puedo resistirme a la tentación de mencionar uno de los dos nombres que firman ese prólogo: Josu Iñaki Erkoreka. Josu Erkoreka es el actual portavoz del PNV en el Parlamento de Madrid, persona que aparece constantemente en los medios de comunicación, y fue director del IVAP de 1995 a 2000; antes había sido secretario de ese instituto y como tal firma el prólogo del libro que ayudó a sacar a la administración vasca de las tinieblas antibotijistas. Recordémoslo cuando le veamos en el telediario.

El Eustat ha recorrido un largo camino desde su –lingüísticamente hablando– pecaminosa creación. Echémosle una hojeda.

“El Eustat es modélico, fíjate, todo se publica en las dos lenguas.” Esta frase se la he oído este mismo año a un compañero de trabajo y refleja una forma de pensar bastante extendida. Basta con tener un buen traductor, y listo. Con esa premisa, el Eustat también podría ser modélico en castellano-japonés.

De las 104 operaciones estadísticas que realiza el Eustat, ¿cuántas han sido pensadas, producidas, gestionadas en euskera?

Por supuesto, hay vasco-hablantes en el instituto, gente que habla euskera entre sí, pero en el trabajo cotidiano el castellano es amo y señor.

No es, evidentemente, una excepción, sino, como no podía ser de otra manera, reflejo de la sociedad en la que se desenvuelve: según las estadísticas (¡LAS ESTADÍSTICAS!) hay en la comunidad autónoma vasca 557.700 vasco-hablantes de más de dieciséis años, un 30,01%, porcentaje que baja al 25,7% si contabilizamos a los habitantes de todos los territorios (sumamos Navarra y las tres provincias vasco-francesas) con población vasco-hablante.

La proporción de uso del euskera es muy inferior a su supuesto conocimiento. En 2007 un 13,89% de los usuarios de la web de Eustat lo hicieron accediendo a la versión (traducida, por supuesto) en euskera. Este porcentaje sí nos refleja la realidad pura y dura.

Hablemos de mi trabajo cotidiano, veamos qué recursos tengo para mi labor cotidiana de traductor, de qué ayudas dispongo:

1. El programa TRADOS. No me voy a extender sobre él, pero, desde luego, podría decirle lo de la canción: “... sin ti no soy nada...”
2. Euskalterm: es el banco terminológico público, gestionado y ofrecido por el Gobierno Vasco. UZEI es el autor del banco Euskalterm, una organización privada, muy conocida por su labor terminológica y lexicográfica. El banco de datos fue propiedad de UZEI hasta que hace ya bastantes años llegó a un acuerdo con el gobierno vasco. Lo que aparece en Euskalterm pasa antes por la Comisión de terminología, donde están representadas la Viceconsejería de Política Lingüística, el IVAP, la UPV (Universidad del País Vasco), Euskaltzaindia (Academia de la lengua vasca), la Consejería de Educación, etc. A Euskalterm se recurre constantemente, muy a menudo en el trabajo cotidiano.

3. ELET (Euskeraz Lan Egiteko Tresnak, herramientas para trabajar en euskera): dentro de la web del IVAP tenemos ELET, en donde encontramos textos legales traducidos, ejemplos de utilización de esta palabra o aquella expresión, diccionarios, gramáticas, modelos estandarizados de documentos...
4. Diccionario Estadístico del Eustat: Éste, y no otro, era teóricamente el tema sobre el que tenía yo que hablar en esta ponencia, pero hay razones que explican esta especie de arrinconamiento; el glosario del Eustat, es, metafóricamente hablando, un infante durmiente. Realizado por UZEL —entidad que ya hemos mencionado— por encargo de Eustat hace unos cuatro años, este glosario se ha quedado detenido en el tiempo, no se actualiza, toda la terminología que aparece en él está recogida en el banco Euskalterm, mucho más amplio y puesto al día constantemente. Podemos decir, entonces, que el Diccionario Estadístico del Eustat “vive” ahora en una casa mucho más grande. A decir verdad, yo ignoraba incluso su existencia “independiente”, nunca me hizo falta, al estar sus términos, como ya he dicho, recogidos en Euskalterm. A pesar de los pesares, el glosario del Eustat fue el primer diccionario en el que uno pudo encontrar términos estadísticos en euskera y sigue siendo una referencia fundamental.

Así como hemos hablado un poco de Historia, miremos un instante al futuro. Como persona que en su vida laboral se dedica a traducir al euskera lo que otros redactan en castellano, quisiera que esa “servidumbre” terminara, que el Eustat y la administración vasca en general fueran capaces de trabajar en euskera (también) directamente, que el supuesto amor a la lengua que tanta gente declara se viera en los hechos, que traductores como yo seamos cada vez menos necesarios.

La administración vasca se encuentra ahora inmersa en el Plan de Normalización del uso del euskera 2008-2012, en el que se hace especial hincapié en que el conocimiento se use, poco o mucho; en que los trabajadores realicen en euskera una parte (que puede ser pequeñísima, para empezar) de sus quehaceres cotidianos, pero esto plantea nuevos y probablemente, eternos, quebraderos de cabeza. Por ejemplo, supongamos que yo soy un funcionario vasco que realiza su labor cotidiana en castellano; animado por el Plan de Normalización del uso del euskera 2008-2012 me decido a realizar la labor X en euskera, lo hago y me sale estupendamente. Muy contento, se lo enseño al jefe, que me dice que muy bien, pero que dónde está la labor X en castellano, que vivimos en una sociedad bilingüe. Le digo que si hago la labor X en las dos lenguas, tardaré lógicamente más y tendré menos tiempo para realizar la labor Y; me dice que yo mismo, pero que todas las semanas, los jueves a las 14:00, tienen que estar listas las labores X e Y en los dos idiomas; yo le mando, al traductor al que toda la vida le he mandado la labor X en castellano para que la tradujera al euskera, la labor X en euskera para que la traduzca al castellano, y él me pregunta cuáles son las grandes ventajas del Plan de Normalización del uso del euskera 2008-2012. Y todo esto sin olvidar que cuando todas las provincias vascas estén unidas políticamente tendremos una sociedad trilingüe en la que la labor X y la Y tendrán que estar todos los jueves a las 14:00 en euskera, castellano y francés sobre la mesa del jefe; y bueno, del inglés y del imparable ascenso del chino mandarín mejor no hablamos.

Cuestiones de difusión estadística relativas al gallego

Carlos L. Iglesias Patiño y M. Esther López Vizcaíno
Instituto Galego de Estatística

Introducción

Actualmente la estadística oficial constituye un elemento indispensable del sistema de información de una sociedad democrática, que proporciona a las instituciones oficiales de la nación, al sector económico y al público, datos acerca de la situación económica, socio-demográfica y medioambiental. Con este fin los organismos responsables de la estadística oficial deben facilitarla con imparcialidad para que los ciudadanos puedan ejercer su derecho de acceso a la información pública, como se recoge en el primero de los Principios fundamentales de la estadística oficial adoptados en la 28ª sesión de la Comisión Estadística de las Naciones Unidas.

No se puede concebir actualmente las estadísticas oficiales sin que exista su difusión, esto es, debe existir una comunicación fundamentalmente unidireccional donde un único emisor transmite mensajes a múltiples receptores mediante un soporte espacio-temporal a través de un canal, utilizando un código común.

En el caso del Instituto Galego de Estatística (IGE), el código común es el típico de la estadística oficial: tablas, diagramas, notación matemática y económica y, por encima de todo, la lengua gallega.

En un momento histórico en que la transparencia de algunos mercados está poniéndose en tela de juicio no conviene perder de vista que este tema de la difusión estadística está comprendido dentro de un campo más amplio de importancia estratégica como es el de la política general de información y está íntimamente relacionado tanto con la regulación –autorregulación o intervención– económica como con la sociedad del conocimiento, de candente actualidad.

A lo largo de sus más de dos décadas de existencia, el IGE ha trabajado en la dirección de convertirse en una de las fuentes fundamentales de obtención de datos de carácter socioeconómico de la Comunidad Autónoma de Galicia (CAG) y de sus principales divisiones territoriales. Es el encargado de recoger, analizar y difundir la información estadística sobre todos los aspectos de la sociedad y economía gallega que le sean encomendados por los programas estadísticos anuales que desarrollan el Plan Gallego de Estadística.

La información estadística tal y como se elabora en el IGE y en las consellerías, tiene las siguientes fuentes diferenciadas: encuestas, censos y registros administrativos. Los temas objeto de estudio se concentran principalmente en dos grandes áreas:

- Área económica: empresas, cuentas, coyuntura...
- Área sociodemográfica: población, mercado de trabajo, condiciones de vida...

Se presentará a continuación el entorno tanto social como institucional, en el cual viene desarrollando el IGE su actividad, en lo que respecta a la lengua.

Entorno social

El grupo de personas que prácticamente no emplean o no tienen contacto con el gallego es pequeño y se reduce progresivamente. El grupo de personas que realizan su vida sólo en gallego también es reducido aunque se incrementa lenta pero paulatinamente.

La inmensa mayoría de la población de la CAG emplea o tiene un contacto significativo con el gallego en su vida cotidiana (enseñanza, mercado, entorno doméstico...). El tiempo o las funciones en que lo emplea es mucho más difícil de medir. En esta elección hay una decisión personal pero también existen unos condicionamientos sociales importantes.

El status del gallego es el de una lengua recién normativizada pero que no está completamente normalizada. Las Normas ortográficas e morfológicas do idioma galego aprobadas en julio de 2003 reformaron las existentes, introducidas por el Decreto 173/1982, de 17 de noviembre, sobre la normativización de la lengua gallega. Basta decir que el Estatuto de Autonomía no se redactó en gallego normativo.

Esta normativa es una construcción supradialectal, que no corresponde con el habla de ninguna zona geográfica concreta y utiliza una ortografía basada en la castellana. La poca distancia lingüística con la lengua oficial del estado, aún minorada por la elección normativa inicial, favorece la práctica inexistencia de conflicto lingüístico en la CAG.

No conviene olvidar que a principios de los 70 del siglo pasado sólo había universidad en Santiago de Compostela. En aquellos momentos esta ciudad no suponía ni el 5 % de la población de Galicia. Además hasta los 90 no se dispuso de la mayoría de las ingenierías, ni de otras titulaciones como veterinaria, biblioteconomía o traducción. Por lo tanto y a modo de ejemplo, los ingenieros o bien se formaban fuera de Galicia o bien no eran gallegos.

Todo lo anterior influye en que la lengua esté poco normalizada en el ámbito tecnocientífico. Por ejemplo, el ensayo y la divulgación científicos, por ahora, son menos abundantes de lo que sería deseable y los libros de texto de la enseñanza no universitaria en numerosas ocasiones son meras traducciones.

Con respecto a su corpus hubo avances significativos en los diferentes campos. En el caso de la estadística también, pero aún queda tarea por hacer tanto de modo puntual como recopilatorio.

A todo lo anterior habría que añadir cierta indiferencia ante la lengua, propia y ajena. Como muestra de lo anterior véase el siguiente titular reciente de La Voz de Galicia: «A valoración social da lingua pexa os avances nas aulas». En este artículo se indica también que la enseñanza «prima os contidos filolóxicos e gramaticais por riba das capacidades comunicativas»; quizás sea una de las explicaciones de este fenómeno.

Otro ejemplo, se encuentra en la onomástica, entre las recién nacidas en el 2006 hay que irse hasta el rango 9º para encontrar un nombre típico gallego “Antía”, o en el caso de los niños al 10º “Brais”. Aunque en puestos anteriores hay nombres con tradición gallega como Alba o Adrián, también conviene resaltar que “Pablo” y “Alejandro”, que en gallego se corresponden con “Paulo” y “Alexandre” respectivamente, son el primero y el cuarto preferido y suponen entre ambos casi el 7%.

En lo que respecta a los instrumentos, no existe problema con los diccionarios generales. También se disponen de bastantes especializados aunque no genuinos de estadística. En lo que se refiere a las nuevas tecnologías, la situación es mejor, existen correctores y traductores de gallego, incluso software libre en gallego.

De un modo aventurado se puede decir que se está posiblemente en un punto culminante en la situación de la lengua gallega. Estamos en el mejor momento de su escritura, pero quizás en la peor situación oral de su historia. Concuerta, en parte, con el proceso de normalización/normativización que ha dejado en una situación de cierto abandono la fonética y la fraseología.

Entorno institucional

La Lei 3/1983, del 15 de junio, de normalización lingüística publicada en el Diario Oficial de Galicia (DOG) núm. 84, del 14 de julio, y en el BOE núm. 213, de 6 de septiembre, y las disposiciones reglamentarias que la desarrollan conforman el marco normativo.

Entre estos reglamentos conviene citar el Decreto 258/2006, de 28 de diciembre, por el que se crea la Comisión Interdepartamental da Xunta de Galicia para la Normalización Lingüística porque en el se regulan las comisiones de trabajo. El IGE forma parte de la Comisión de Traballo de Normalización da Consellería de Economía e Facenda.

Y también el Decreto 132/1984 por el que se establece el procedimiento para la fijación o recuperación de la toponimia de Galicia, así como el Decreto 174/1998 que regula las funciones y composición de la Comisión de Toponimia, dado que el IGE ha formado parte de ella.

Otro elemento relacionado con el gallego y la difusión estadística, aunque de modo indirecto, ha sido la legislación sectorial, donde además de la propia Lei de Estatística de Galicia están las leyes de los tres planes y los decretos de programación anual. En los dos primeros planes existía una identificación implícita entre el Plan Gallego de Estadística (PGE) y el inventario, ya que las diferentes actividades que se incluían en el programa se seleccionaban entre las que conformaban el plan.

En el nuevo marco del tercero, un plan redactado en forma de metas que se desarrollan en cada programa estadístico anual (PEA), el inventario de actividades desempeña un papel de mayor importancia, por eso está actualmente en la web del IGE.

Dentro de este punto en lo relativo a la normativa estadística más técnica merece la pena mencionar el asunto de los proyectos. Toda actividad estadística en la que la administración gallega participe en alguna fase diferente de la mera difusión necesita disponer de un proyecto técnico. A este tipo de actividades se las denomina operaciones estadísticas.

Para su redacción se confeccionaron unas pautas que dependen del tipo de operación. Estas están redactadas en gallego, como no podía ser de otro modo. Los proyectos y los borradores, de los cuales una parte importante corresponden a las Consellerías y sus organismos dependientes, se realizan también en gallego aunque su corrección lingüística es manifiestamente mejorable.

En cuanto a las incursiones dentro de la jerga más técnica del gallego estadístico es interesante destacar, por una parte, los diferentes temarios de oposiciones para las escalas de estadísticos de la CAG; por otra, los artículos en las publicaciones periódicas *Boletín de Actualidade Estadística* y *Boletín de Series Estadísticas de Galicia*. En este último incluso se puede hablar de un “exceso de celo” porque se traducían algunas contribuciones redactadas originariamente en castellano.

Todo lo anterior sin que hayan existido nunca lingüistas profesionales en la plantilla del IGE, por lo que la revisión del gallego, en algunos casos, se ha realizado por agentes externos: Secretaría Xeral de Política Lingüística (Presidencia da Xunta de Galicia) y el Servizo de Tradución (Consellería de Presidencia, AA.PP. e Xustiza). Esta revisión se centra en la corrección pero no en el estilo ni en la oportunidad, como es lógico.

Además la CAG dispone de manuales de lenguaje administrativo y de documentación administrativa general y organiza regularmente cursos sobre la materia. Sin embargo, aunque existen diccionarios, léxicos y manuales para el lenguaje económico y otras jergas técnicas, hasta donde se tiene conocimiento, no se han realizado cursos de lenguaje socioeconómico que pudieran mejorar la corrección y el estilo del gallego en la difusión estadística.

Visto el entorno en que desenvuelve sus funciones el IGE, se considerará ahora la función de difusión en relación con la lengua gallega de modo más directo, que se clasificará en: la de tipo

tradicional, la que se realiza a través de la web y la que denominaremos relacional. Esta última suele tratarse poco en la literatura pero en este campo tiene mucha importancia como se verá posteriormente.

Difusión desde una visión tradicional

El IGE inicialmente estaba a la vanguardia y colaboró decisivamente en la puesta en marcha del proceso de normalización del gallego, por ejemplo, el anuario *Galicia en cifras*, cuyas primeras ediciones llegaban a casi todas las bibliotecas públicas o escolares, contribuyó muchísimo porque en aquellos momentos existían pocas obras con un contenido técnico tan variado en gallego normativo (su primera edición estaba en la calle sólo dos años después de la difusión de la normativa de la lengua gallega y en aquellos momentos había órganos de la CAG que aún editaban en castellano o en bilingüe). Esta labor fue continuada por el *Datos estadísticos básicos*.

Otro ejemplo fue la traducción del TARIC que junto con la posterior de la CNAE-93, constituyeron un hito en este proceso de normalización. La primera supuso un esfuerzo enorme por parte del Servizo de Tradución da Consellería da Presidencia que se cristalizó en la publicación del IGE, dentro de su antigua colección de Comercio exterior, *Nomenclatura do Arancel Integrado Comunitario detalle a catro algarismos*; que se vería continuada mediante las publicaciones de la CNAE, una con la estructura y otra con las notas explicativas.

Conviene resaltar, por su especificidad, la labor del IGE en el campo de la difusión documental donde merece la pena destacar dos publicaciones: *Bibliografía de fontes estatísticas 2002* y *Aproximación ás fontes e conceptos de información estatística no contexto socioeconómico galego*.

En la actualidad, los productos de difusión, son habitualmente resultado de las operaciones y actividades estadísticas. En este sentido la difusión se materializa a través de los siguientes medios:

- La web es el principal medio canalizador de la difusión del Instituto (<http://www.ige.eu>). En ella se pueden encontrar las metodologías y los cuestionarios de sus operaciones estadísticas, clasificaciones estadísticas, calendario y catálogo de publicaciones, folletos en red, tablas estáticas y dinámicas...
- Folletos impresos gratuitos de carácter específico
- Publicaciones de carácter general: *Galicia en cifras*, *Datos estadísticos básicos de Galicia*, el futuro Anuario estadístico municipal...
- Notas de prensa para los medios de comunicación

La lengua que se utiliza habitualmente en la difusión del IGE es el gallego. La página web está en gallego y castellano, aunque las tablas estadísticas están únicamente en gallego. En el resto de los medios utilizados para la difusión (folletos, publicaciones y notas de prensa) se utiliza como lengua casi única el gallego.

Todas las publicaciones, ya sean de carácter especializado o general, se elaboran en el propio Instituto y una vez terminadas se envían a los técnicos de la Secretaría Xeral de Política Lingüística para que revisen el gallego adaptándolo a la normativa vigente. Para asegurar la corrección de la lengua en la información que está a disposición del público en la web, se utilizan los recursos ya citados que ofrece la Xunta de Galicia, como traductores, correctores, diccionarios y vocabulario existentes, en formato impreso y en la red.

En cuanto al procedimiento que sigue el IGE para la elaboración de clasificaciones estadísticas, que permiten la comparación de los resultados estadísticos con los de otros ámbitos temporales y

espaciales en el marco de la estadística oficial, señalar que se traducen dentro del propio Instituto y a continuación se dirige al Servicio de Traducción de la Xunta de Galicia para que sólo realice la revisión del gallego, teniendo en cuenta su experiencia por la participación en la traducción de versiones anteriores o de otras clasificaciones.

En lo que hace referencia a los cuestionarios que se utilizan en el IGE para recoger la información estadística que aportan los hogares y las empresas, se utiliza la lengua gallega, excepto en los dirigidos a empresas que tienen sede social fuera de la CAG en cuyo caso se le envía los cuestionarios en castellano. Tanto la elaboración como la revisión del gallego se llevan a cabo dentro del Instituto, empleando los recursos mencionados con anterioridad.

Conscientes de la importancia que, para una oficina estadística, tiene el hecho de remitir la información a la prensa de manera directa y clara, se creó un apartado en la página web del IGE dedicado a los medios de comunicación (PRENSA). Este apartado aglutina las notas que se envían a los medios de comunicación. Para su redacción hay un protocolo en el que se establece que sean claras, precisas y sencillas ('plain language', sin lenguaje técnico) y escritas íntegramente en gallego. La revisión de la lengua en las notas de prensa se lleva a cabo dentro del Instituto.

Difusión vía web

La difusión de información estadística por vía telemática se remonta en el IGE a finales de 1993 en que se inició un proyecto de servidor videotex acogido al programa ARCO gestionado por la Dirección General de Telecomunicaciones del Ministerio de Obras Públicas y Transporte.

La explosión de Internet provocó el abandono de esa plataforma dado el carácter de servicio sustitutivo, de tal modo que el Instituto con el tiempo tuvo que acabar por cerrarlo y quedarse sólo con la web por razones de economía.

Como consecuencia de la importancia que ha cobrado Internet en la difusión estadística, se ha convertido en la actualidad en el principal mecanismo de difusión del IGE, se vio la necesidad de ampliar el sistema de metainformación que facilitase a todos los usuarios de la web el acceso, consulta, comprensión e interpretación de la información en ella contenida.

La implantación en la web del banco de datos trajo consigo la necesidad de crear una base de datos de definiciones relacionadas con las estadísticas elaboradas o difundidas por el Instituto. Estas definiciones están escritas únicamente en gallego. La motivación fundamental fue dar la posibilidad al usuario, cada vez que se dispusiese a crear una tabla dinámica, de consultar de modo inmediato las definiciones de las variables de dicha tabla y así obtener ayuda para interpretar la información.

Al crear la base de datos de definiciones se tuvo en cuenta el objetivo de lograr la consistencia de la información a lo largo de toda la web, es decir, la existencia de una única base de datos, y sobre todo, la existencia de un único concepto.

Algunas de las principales ventajas de tener las definiciones organizadas en una base de datos son la integridad y la autonomía de gestión, puesto que cualquier modificación introducida en las definiciones se reflejaría de modo automático en la web. Una vez introducidas las definiciones en la base de datos, se va a crear un apartado que permita el acceso a las mismas por un simple orden alfabético o mediante un acceso temático. Además en esta aplicación se podrán realizar búsquedas analíticas.

Otra labor importante, que se llevó a cabo en la página web del IGE relacionada con el gallego, es la adaptación de la relación detallada de entidades y núcleos de población existentes en cada municipio de Galicia, en el Nomenclátor elaborado por el Instituto Nacional de Estadística, a la

legislación vigente en materia de toponimia en la CAG.

La Ley 3/1983 de normalización lingüística dispone en su artículo 10 que los topónimos de Galicia tendrán como única forma oficial la gallega. Además, este mismo artículo establece que le corresponde a la Xunta de Galicia la determinación de los nombres oficiales de los municipios, de los territorios, de los núcleos de población, de las vías de comunicación interurbanas y de los topónimos de Galicia.

En este sentido, cada vez que el INE publica su Nomenclátor, el IGE lleva a cabo la corrección de los topónimos teniendo en cuenta los establecidos en la Toponimia Oficial de Galicia y a continuación se publican en la página web.

Dentro de la línea de adaptación de la información de la web al perfil del usuario final, este año se puso a disposición de los centros educativos de ESO, Bacharelato y ciclos formativos un portal educativo escrito únicamente en gallego.

Además del objetivo de adecuarlo a las necesidades de alumnos y profesores, el desarrollo del portal educativo responde, fundamentalmente, a potenciar el uso de la estadística pública para que se conozcan y aprovechen las posibilidades que ofrecen sus productos y actividades.

Se persigue, de este modo, que los jóvenes conozcan los datos y las fuentes estadísticas disponibles y que se acostumbren a analizarlos e interpretarlos, para que puedan disponer de más habilidades y destrezas que les permitan emplear este recurso en su etapa adulta y en su vida laboral; es decir, se trata de que el portal educativo constituya una herramienta básica para la formación de futuros usuarios.

En este portal se ofrece a los profesores material didáctico actualizado, adaptado al entorno en el que desarrollan su labor educativa y que, además, favorezca el análisis y la interpretación de los datos y la información. Todo ello en la lengua propia de Galicia, en la que deben impartir sus clases según el Decreto 124/2007, de 28 de junio.

Difusión relacional

Se denominará relacional a un tipo de difusión, en cierto modo indirecta, entre la estándar y la a medida, que se basa en la colaboración institucional. Aunque no ha estado lo suficientemente formalizada en algunas ocasiones, ha sido fructífera en sus resultados, por lo que tuvo y cada vez tiene más importancia.

Como ejemplos pasados, y sin ánimo de exhaustividad, se puede mencionar la elaboración, en colaboración con otros departamentos de la Xunta de Galicia, de la publicación seriada *Galicia* que ofrecía una imagen global de su realidad económica, social y cultural. O también la aportación al *Manual básico de economía en galego* de La Voz de Galicia, una de las obras primeras y de mayor difusión sobre conceptos económicos.

En la actualidad también se colabora con el Consello da Cultura Galega, apoyándolo en la difusión de estadísticas. Como ejemplos su *Anuario de Estatísticas Culturais* o *Ocio e hábitos culturais dos galegos*, esta última, explotación y difusión del módulo correspondiente de la Enquisa de Condicións de Vida, recogida y procesada por el IGE hasta la disposición del fichero final anonimizado. Ambas publicaciones fueron elaboradas por el Observatorio da Cultura Galega del Consello.

Tampoco conviene olvidar la colaboración con la Fundación Caixa Galicia con la que además, recientemente, se concedió el Premio Galicia de Estadística, del que se editó un folleto donde se

recogían los discursos del evento que salvo el ganador, por razones obvias, están en gallego.

En el tema de colaboraciones con la administración local merece mención especial el convenio de colaboración desde el año 2001 entre el IGE y el Concello de Santiago de Compostela. Tiene por objeto el establecimiento de una colaboración entre las dos instituciones a través del apoyo técnico, organizativo y documental.

Fruto de esta colaboración y de sus propios datos, el ayuntamiento de Santiago publica un anuario municipal en gallego, en el que presenta un conjunto de datos cuantitativos que perfilan una fotografía del presente compostelano y, con datos de años precedentes, tratan de identificar tendencias de lo que vendrá en los próximos años.

En este tipo de difusión, basada en las relaciones del IGE con la empresa/institución, directamente o a través de sus funcionarios, conviene destacar las antiguas con la Direcção Regional do Norte do Instituto Nacional de Estatística portugués por su intensidad y naturaleza.

A modo de ejemplos paradigmáticos de esta difusión cooperativa podemos citar el *Anuario Estadístico Galicia-Norte de Portugal* que en algunas ediciones fue bilingüe (gallego y portugués) e incluyó pequeños textos de introducción a los lectores en los diferentes capítulos. O el *Comercio Intracomunitario da Eurorrexión Galicia-Norte de Portugal 1995-1997* en dos ediciones con idéntico contenido (una en gallego y otra en portugués) que disponía de mucha más información textual que numérica.

Esta colaboración redundó en la mejora del tecnolecto socioeconómico al confrontarlo con su lengua hermana, miembros ambas del romance ibérico occidental, además de suponer una indudable proyección a nivel europeo de la lengua propia del noroeste peninsular.

También conviene apuntar que el IGE es socio institucional de la Sociedade Galega para a Promoción da Estatística e Investigación de Operacións (SGAPEIO) que tiene como objetivo principal el fomento y divulgación de las técnicas y aplicaciones de la estadística y de la I.O. en el ámbito gallego.

Además de la celebración de sus congresos bienales y cursos, conviene citar en el ámbito de la difusión estadística, por una parte, su boletín *Informest* donde se presentan artículos de carácter divulgativo y novedades del IGE, entre otras noticias, y por otra, el Premio IGE de Innovación Pedagógica.

En esta misma línea, donde hay amplio campo para la mejora, no debemos olvidar la labor docente del IGE en sus cursos de estadística aplicada a la administración y a la enseñanza media y superior dentro de los programas de formación permanente de la Escola Galega de Administración Pública o de la Consellería de Educación e Ordenación Universitaria o bien de las Universidades gallegas.

Conclusión

Resumiendo, el IGE ha sido reactivo y catalizador importante en la combinación lengua gallega y ciencias sociales en su vertiente cuantitativa, tanto en lo relativo a la información como incluso al conocimiento estadístico, pero de un modo voluntarista y quizás con poco rigor lingüístico. Han predominado las soluciones ad hoc típicas de la aplicación del dicho: lo urgente impide lo importante.

Actualmente, el Instituto se está quedando en un segundo plano porque se han presentado otras demandas sociales más acuciantes que la estrictamente lingüística, por ejemplo, profundizar en el estudio del mercado de trabajo; o políticas, informar sobre las condiciones de vida incluyendo ingresos familiares y el propio uso y conocimiento del gallego.

Es decir, ha habido cierto interés cuantitativo por el gallego que no se traduce en un interés cualita-

tivo, de perfeccionamiento de la lengua; priman otras cuestiones. Además existe una preocupación creciente por la información estadística producida en la propia CAG que repercute en el tratamiento de la metainformación, dada la escasez de recursos.

Se considera que en lo que falta por realizar en el corpus estadístico de la lengua gallega, el IGE debe tener un papel preponderante y con este escenario se trabajará tanto en la difusión estadística como en la necesaria coordinación. La meta es conseguir un lenguaje cada vez más controlado que abandone las imprecisiones y las ambigüedades.

Una vez que se haya conseguido disponer de proyecto para cada operación estadística se realizará un esfuerzo en la precisión lingüística y con el tiempo se intentarán difundirlos siguiendo el precedente de la *Documentación técnica de operaciones estadísticas*, aunque probablemente se tenderá a emplear medios informáticos.

A corto plazo, se prevé ampliar el número de variables que se presentan en el Inventario dentro de la web y difundir, en breve, las notas explicativas de la CNAE-2009 traducidas al gallego. A medio, compilar y divulgar un vocabulario estadístico al menos trilingüe y, a largo plazo, convertir en un diccionario estadístico las definiciones. En la actualidad, están orientadas a facilitar la comprensión de las tablas dinámicas de la web por parte del usuario, pero no constituyen un “verdadero” diccionario de conceptos y variables.

Para finalizar se puede decir, parafraseando la exposición de motivos de la Lei 3/1983 de normalización lingüística, que el gallego perdió la batalla de la imprenta, pero está ganando la de las tecnologías de la información y comunicación (TIC); mientras que el IGE, en el tema que estamos tratando, ganó la batalla de la imprenta si bien está perdiendo temporalmente la de las TIC, aunque se tiene la esperanza de que no se pierda la guerra.

La lectura de les dades estadístiques en els mitjans de comunicació¹⁰

Albert Closas

Director del programa Valor Afegit de TVC

Per què ens interessa l'estadística?

L'estadística té un valor incalculable en l'àmbit dels mitjans de comunicació perquè, a més de ser un reflex de la societat, és una arma que permet al periodista una gran llibertat si és capaç d'interpretar les dades correctament. És a dir, que tot i que hi hagi una voluntat de manipulació per part d'un sector de la societat, les dades són les que són, i això permet jugar amb avantatge i, si cal, contradir aquest sector i fer veure a la població la situació real.

Què ens interessa de l'estadística i què ens hauria d'interessar més?

Ens interessen i ens han d'interessar els diferents tipus d'indicadors: els econòmics, els socials i els de població, com també tot el que pot anar més a l'alça o a la baixa en funció dels mitjans de comunicació i de la percepció de la realitat (els anomenats *in* i *out*). El número u per excel·lència des que van començar les polítiques monetàries als anys vuitanta són els preus, tot i que sovint es tendeix a buscar l'indicador més que no pas aprofundir en la informació estadística, perquè els grans números sempre impressionen més. La repercussió real de l'IPC, per exemple, és força diferent del que s'anuncia als mitjans. Per molt que es digui que l'ha fet pujar segons quin producte, sabem que n'ha estat un altre el que ha acabat per determinar-ne la pujada.

Pel que fa a l'*in* i l'*out*, crec que la moda de l'IPC anirà a la baixa, perquè cada vegada tendim més a la deflació i, per tant, ja no es parlarà de la inflació.

El que sí que ens hauria d'interessar més és la inflació d'actius, que és la gran oblidada per les institucions d'estadística de prestigi, però que els bancs han seguit al peu de la lletra, cosa que ha provocat la crisi financera. El terme *inflació d'actius* fa referència tant als actius immobiliaris com als financers. D'una banda, les fonts dels actius immobiliaris són moltes i estan molt disperses i, de l'altra, encara no és del tot clar què s'entén per actius financers. Si fóssim capaços de tenir en compte tot un conjunt d'indicadors alhora (com ara els preus en borsa, l'evolució dels productes financers, etc.), això ens hauria de permetre detectar les bombolles immobiliàries i també mesurar-les, cosa que ara per ara ens és impossible.

Uns altres indicadors són l'atur i l'ocupació, que són valors que aniran a l'alça perquè afecten directament la vida de la població. Des dels mitjans s'hauria d'intentar incidir més en estadístiques

¹⁰ Aquesta ponència és una transcripció de l'enregistrament que es va fer durant les Jornades, degudament autoritzada per l'autor.

qualitatives i menys en la precarietat, que ja hi queda molt reflectida: el problema actual no és l'atur i les conseqüències que comporta, sinó el gruix de població que treballa en feines mal remunerades i que no arriba a final de mes.

El producte interior brut és un valor a l'alça, però té l'inconvenient que no és actual sinó que es demora uns quants dies. El gran perill és que ens el mirem més com a indicador que com a estadística, sobretot per l'afany dels últims temps de demostrar que més que en recessió i crisi, som a tocar de la depressió. Un exemple clar és el que va passar amb l'INE, que va fer un avançament de la comptabilitat nacional trimestral el catorze de novembre, en què anunciava una caiguda del PIB intertrimestral, cosa que va provocar una psicosi general a la població, que al seu torn va provocar una caiguda del consum i de la inversió, entre altres.

També ens hauria d'interessar més un altre valor que va a l'alça: el dels salaris. No és que no hi hagin dades sobre els salaris, però haurien de ser més detallades i haurien d'incloure informació qualitativa, perquè actualment encara són massa disperses, massa generals, massa poc representatives i de difícil accés, i la població estaria molt més receptiva si s'hi veiés reflectida. Hauria de ser informació més fragmentada, en cinques de població, de manera que es veiés realment el percentatge de població més rica i més pobre, i les desigualtats salarials.

Dos altres indicadors que ens haurien d'interessar més són l'evolució de l'endeutament de les famílies i de les empreses i l'evolució de les balances exteriors. El primer està relacionat amb la inflació d'actius i, per tant també amb l'evolució de l'estalvi. El segon apareix ara a tots els mitjans i és el que no ens hem atrevit mai a explicar en profunditat. A banda d'aquests dos, hi ha les grans assignatures pendents de l'economia espanyola: la balança per compte corrent, la balança de pagaments i la balança de capital. Espanya és, junt amb Grècia, el líder mundial del món industrialitzat del dèficit per compte corrent. Tots aquests indicadors, igual que la inversió en capital humà, són bàsics, però no criden l'atenció precisament perquè no són immediats, tot i que estan a la base de la solidesa d'un país i de la seva economia. És per això que calen més estadístiques d'inversió i d'innovació, amb indicadors més populars, com ara estadístiques en ensenyament, titulats universitaris, oferta i demanda, nombre de patents, etc., per arreglar aquests problemes estructurals.

Pel que fa als indicadors socials, em sembla que el problema més important és la falta de formació dels periodistes no especialitzats en economia. Per tant, segmentar la informació facilitaria la interpretació de les dades i també permetria descobrir més detalls, perquè les dades ja són al carrer, només cal saber-les interpretar. Un exemple que no hi ha formació estadística és que, en una afirmació com ara que les dones centenàries quintupliquen els homes centenaris, la població no es fa la idea exacta del concepte "cinc vegades més". (Segons això, Matussalem devia ser una dona, oi?).

Hi ha molts fenòmens socials interessants, però en comentó només quatre:

En primer lloc, el fenomen de la generació Franco: la generació més nombrosa per any de naixement és la del 1975, la de després de Franco.

En segon lloc, el nombre d'andalusos residents a Catalunya s'ha reduït notablement des de l'any 2000, i tot i que la raó sigui vegetativa, una disminució de 10.000 per any és considerable.

En tercer lloc, en el segment de gent de 55 a 64 anys és més nombrós el grup de gent nascuda fora de Catalunya que el grup de nascuts a Catalunya.

En quart lloc, tot i que es creu que el nombre de dones llatinoamericanes a Catalunya és superior que el d'homes, s'està observant els últims anys que comencen a igualar-se.

Com ens interessa l'estadística?

Per a la població en general i sobretot per als treballadors dels mitjans de comunicació és important que l'accés a la informació estadística sigui ràpida, d'interpretació clara, ben segmentada i sobretot que s'actualitzi sistemàticament (per exemple, mitjançant un calendari setmanal). Seria important també l'accés als aclariments i a les interpretacions dels tècnics, la comparació de dades amb Europa o amb els Estats Units i caldria ser una mica més proactius.

Quant als llocs web de les institucions estadístiques, el de l'Idescat és el que ha fet un salt qualitatiu més gran els últims anys: ara és més visible, més clar i ordenat i disposa de connectivitat entre diversos indicadors, cosa que s'agraeix moltíssim.

Vicis i defectes del llenguatge estadístic en els mèdia¹¹

Josep Maria Ureta

Periodista a la secció d'opinió d'El Periódico

Malgrat que hi hagi la creença que les ciències es divideixen en dos, jo crec que es divideixen en tres: les naturals, les humanístiques i les recreatives, que són aquelles a què ens dediquem els periodistes. És tot el que fem per poder explicar, i que no forma part de cap de les ciències. Avui em toca de parlar dels “Vicis i defectes en els mèdia”, un títol que m’ha marcat molt perquè m’obliga a parlar malament de la meua feina.

Durant la Transició la informació econòmica era pràcticament inexistent i estava vinculada a la informació política. Els suplementes d’economia que sortien, només de tant en tant, a la premsa, eren elaborats per economistes i no per periodistes. Amb el temps, però, i gràcies a la important feina dels serveis d’estudis de bancs i caixes, que van començar a redactar els butlletins en català, els mitjans de comunicació van poder començar a treballar en termes econòmics, cosa que va ser un gran pas per a la normalització. Cal reconèixer, doncs, el gran esforç i la millora fets per part dels lingüistes, els economistes, els estadistes i els periodistes, perquè si actualment hi ha una informació econòmica sòlida i fiable és gràcies a tots.

Paral·lelament, però, hi ha un sentiment d’insatisfacció per una certa pèrdua de rigor i d’exigència comuna a causa de les noves tecnologies, ja que a vegades ens preocupa més l’estètica (que els gràfics quedin bonics i amb molts colors) que no pas si una dada és correcta o no. Aquesta és la lluita primordial en aquests moments.

Un altre vici dels mèdia són els errors de traducció que es cometen. N’hi ha hagut tants que s’ha pogut fer un diccionari de *falsos amics*. Sovint ens trobem, en textos en llengua catalana, termes com ara *expectativa de vida*, que en català s’ha de dir *esperança de vida*, però com que català i castellà són llengües germanes, fem la traducció automàtica. El problema, doncs, rau en la traducció literal. Ricard Fité, cap de lingüistes d’*El Periódico*, l’únic diari que surt cada dia en dues llengües, defensa que la millor manera de lluitar contra els errors lingüístics és anar contra la inèrcia.

El traductor automàtic d’aquest diari consta, actualment, d’una base de dades de cinc-centes mil paraules (cal remarcar que, a l’inici, en tenia “només” deu mil) i no tradueix les que tenen diversos significats depenent del context per no produir cap error de traducció. En aquest sentit, es va demostrar, en una prova de la Universitat Politècnica de Catalunya, que pel que fa al llenguatge estadístic és més fiable un traductor automàtic que un traductor humà, i em sembla que és pertinent dir-ho precisament en aquestes Jornades, que tracten de llengua i estadística, ja que els índexs de fiabilitat són bàsics i haurien de ser la preocupació principal.

Precisament Joaquín Leguina parla dels índexs de fiabilitat en un article que us recomano, titulat *Estadística y medios de comunicación*, en el qual reflexiona sobre les mostres aleatòries, sobre les que ho són de debò, com ara dades de l’IPC o la inflació, i sobre les que no ho són de debò, tot i

¹¹ Aquesta ponència és una transcripció de l’enregistrament que es va fer durant les Jornades, degudament autoritzada per l’autor.

que es venen com si ho fossin, com ara les enquestes d'opinió.

Cal fer la distinció, doncs, entre què són enquestes socials i què són enquestes econòmiques. Una dada vital que cal tenir en compte és qui hi ha darrere de l'encàrrec de cada enquesta, perquè la fiabilitat de la interpretació d'unes dades sempre depèn de qui les publica. Als Estats Units, per exemple, el nombre de tímids va passar d'un a trenta milions després que una agència de relacions públiques interessada transmetés aquesta dada a una revista. Per tant, utilitzar les dades per a beneficis propis, independentment de la causa, és manipulació, i s'ha d'anar molt amb compte, perquè sovint els periodistes considerem com a objectives estadístiques que només responen a interessos personals. Això passa (tant a periodistes, com a estadístics, com a lingüistes) perquè no estem prou atents.

Un dels vicis dels periodistes, que va lligat, per això, als vicis dels estadístics, és formular preguntes incorrectes, com ara "Quants musulmans hi ha a Catalunya?". Aquest tipus de preguntes no s'haurien de fer, però, si es fan, els tècnics haurien, en primer lloc, de no contestar i, en segon lloc, donar les explicacions pertinents sobre per què no poden respondre.

Per acabar, diré que hauríem d'aprendre a utilitzar més bé l'economia i l'estadística, a interpretar millor les dades i a ser més empírics. I vull remarcar també que tot i que a l'Estat espanyol s'acostumi més a qüestionar-los, els números es calculen, i que si estem immersos en una crisi financera és per una d'aquestes dues possibilitats: o les dades estadístiques eren incorrectes, o ho era la interpretació que se'n va fer.

L'adaptació de les notes de premsa estadístiques en els mèdia

Joan Roy

Responsble de publicacions del Grupo Dossier

Havent intervingut ja els meus companys per exposar aspectes tan importants en la difusió de dades estadístiques socialment i informativament significatives com són els “Vicis i defectes del llenguatge estadístic”, que ha fet l'amic Ureta, o la “Lectura de les dades estadístiques als mèdia”, de l'Albert Closas, em toca a mi parlar d'una qüestió també cabdal com és l'adaptació de les dades als mèdia.

En això, potser caldria tenir primer en compte aquella dita que assevera que “el periodisme ha d'informar sovint de coses que demà seran oblidades, si bé que les dades romanen...”. Aquest, em sembla a mi, és el sentit de l'estadística i la seva grandesa en establir paràmetres reals d'interlocució informativa i, per tant, periodística.

I

Ara com ara l'economia ho determina tot, menys l'amor quan és de debò, que dirien els poetes, però això ja és una altra qüestió.

Als efectes informatius que ens ocupen, la determinació de l'economia arriba, fins i tot a la política, a la cultura i al lleure. Per resumir, podríem dir que l'economia ha esdevingut un terreny comú de preocupació social, que com a disciplina específica arrengrera un seguit de paraules, conceptes i dades d'interès preferent, i que col·loca els professionals dels mitjans de comunicació en primera línia de la consciència i la representació general.

Els periodistes, doncs, estem cada dia més abocats a parlar d'economia i a fer servir i divulgar estadístiques, perquè resulten útils a l'hora de cohesionar el tot social que formen política, economia i societat.

Cal assenyalar en això el triomf en els darrers anys del model anglosaxó de comprensió i identificació de la realitat, pel seu caràcter sempre i prioritàriament matematicoestadístic. Aquest triomf no ha estat només un modisme o conseqüència de la dinamització d'Occident, sinó que ha resultat d'una voluntat manifesta de la comunitat internacional per establir uns criteris de comunicació comuns sobre economia, sobretot, des la creació i evolució dels grans organismes internacionals: ONU, Banc Mundial, Fons Monetari Internacional, i, més tard, també la Unió Europea o l'Organització Mundial del Comerç.

Només cal recordar que a les darreries del segle XX hi hagué una revisió a fons de les nomenclatures estadístiques internacionals, que van comportar el desenvolupament de noves nomenclatures fins a formar un sistema integrat i harmonitzat. Així, l'acord de la Comissió Estadística de les Nacions Unides i de la Conferència d'Estadístics Europeus del 1972-73, per aparaular un sistema de classificacions econòmiques universals, unificant el que fins llavors era divers, tot plantejant serio-

sos problemes de comparació entre països. Així, s'establiren directrius per al disseny i l'elaboració d'un sistema integrat de nomenclatures i classificacions d'activitats i productes (SINAP), i per a la creació de classificacions específiques de productes orientada, prioritàriament, a la computació de la producció i al comerç internacional.

Situats en aquest punt, cal recordar també que justament per això les classificacions econòmiques amb finalitats estadístiques s'agrupen, a grans trets, en dos grans tipus: les d'activitats econòmiques i les de productes.

Les classificacions d'activitats econòmiques apleguen unitats productives segons tipus d'activitat amb vista a la jerarquització d'aspectes relacionats amb el funcionament econòmic: producció, utilització dels factors de producció (treball, capital, materials) o rendibilitat. Les classificacions de productes segueixen relacions funcionals entre les principals operacions de producció, comerç o transport. I en això es distingeixen, d'una banda, els béns transportables o mercaderies i, de l'altra, els béns no transportables o serveis.

També cal saber que s'entén per activitat econòmica tota acció productora resultant d'una concurrència de mitjans o factors que condueix a la creació de béns o serveis, amb independència que les activitats es facin amb finalitats lucratives o sense. Així, el sector públic s'afegí en aquest conjunt de reglamentacions a l'hora d'establir els seus còmputos, adequant-hi el seu ordre impositiu, comptable i pressupostari.

II

A la fi, tot plegat ha conformat una especialització informativa d'abast mundial que sovint contrasta, sobretot a casa nostra, amb l'escassa formació en temes d'economia i d'estadística per part de la classe periodística, més abocada per raons conegudes (democràcia recuperada) a la primacia de la política en la informació.

En tot cas, la dominància de l'economia ha imposat una mena de curs accelerat en macroeconomia i estadística, precisament per bastir la informació corresponent als paràmetres principals de l'evolució social, com ara el creixement, la inflació o l'atur i en detriment d'altres, també compilats, si bé que menys referenciats, com són els preus industrials o el dèficit comercial.

Així mateix, la transparència assolida quant als pressupostos i el finançament de l'Estat han acabat d'arrodonir la presència de l'economia i la macroeconomia en la informació periodística regular. També l'evolució consensuada de la Unió Europea, sobretot des de la creació de l'euro i de la zona euro, ajudà (i molt) a adequar les informacions econòmiques continentals als paràmetres estadístics i conceptuals establerts per la mateixa Unió, i reflectits, consecutivament i temporalment, per organismes com l'Eurostat. No diguem en les actuals circumstàncies de crisi financera mundial, on els conceptes especulatius i borsaris que hi juguen excedeixen, fins i tot, el coneixement de molts especialistes del ram econòmic.

Amb tot, i al marge de crisis d'un o altre tipus, la presència de l'economia i la macroeconomia als mitjans ha registrat una adaptació, podríem dir-ne que tutelada primer per l'existència d'organismes especialitzats de gran abast, com els europeus i els espanyols, i per altres de propis, com l'Idescat, sempre esforçat a dotar la realitat econòmica catalana d'una dimensió nacional i territorial.

Cal dir que en aquest sentit la disponibilitat de productes estadístics ha estat un factor rellevant a l'hora de difondre l'estadística oficial i la metodologia de treball a l'entorn de les classificacions. A casa nostra, doncs, s'ha anat consolidant amb la seva pràctica i s'ha concretat en processos com ara la normalització de la llengua, el procés de socialització de la classificació estadística als agents del Sistema estadístic de Catalunya o l'oficialització que difon les normes de classificació.

III

En el cas català, doncs, la tasca de l'Idescat ha estat titànica, avalada per l'Administració i per la creació d'altres importants organismes d'estipulació econòmica com ara el Consell de Treball Econòmic i Social de Catalunya, per no parlar de la destacada i més antiga tradició dels serveis d'estudis de caixes i bancs, particularment els de la Caixa o Caixa Catalunya.

Tot aquest bagatge és el que explica realment quin ha sigut i en quin estat es troba l'adaptació de les notes de premsa estadístiques als mitjà, si més no perquè sense aquestes, sense la voluntat divulgadora i de transparència dels grans organismes de compilació econòmica del país, hauria sigut molt difícil que els periodistes féssim bé la nostra feina d'informar de les més preocupants qüestions econòmiques.

En aquest sentit, ja fa molts anys que l'excel·lència en l'elaboració i difusió d'aquestes notes ha permès que els mitjà informin amb assiduïtat i tenacitat dels grans temes de preocupació econòmica que faciliten entendre, sempre com un conjunt, l'economia familiar i l'empresarial. Conceptes clàssics ja esmentats com ara creixement, atur, ocupació, inflació, han obert pas a d'altres de més complexos com ara demanda interna, consum, balança de pagaments, productivitat i competitivitat, o bé innovació, recerca i desenvolupament, etc., i tots han estat classificats de forma regular i convenient pels nostres organismes, que permeten un grau de consciència social elevat, reflectit, per exemple, en la negociació de convenis laborals o en la postulació de mesures de correcció o adequació econòmica per part de l'Administració.

El balanç, doncs, és molt positiu, però justament per això planteja nous reptes, la definició dels quals requeriria la participació pràctica i efectiva de la classe periodística per transmetre als organitzadors de dades nacionals els neguits o els buits que es detecten a la societat. Tots plegats ens hem d'implicar en un millorament col·lectiu, per fer que es generalitzin les bones pràctiques estadístiques, la gent se les faci seves i cada cop hi hagi menys intermediaris i filtres periodístics habituals, entre les dades i les valoracions, ambdues imprescindibles per al bon funcionament d'una societat moderna i avançada.

Com presentem la informació estadística a Internet?¹²

Xavier Rius

Director d'e-notícies.com

En general, com que en el món dels mitjans de comunicació el més important és la rapidesa, als periodistes els agrada que les notícies els arribin pràcticament acabades, de manera que només hagin de fer un parell de retocs abans de publicar-les. Entre altres coses, és millor que les notícies siguin simples i ràpides de llegir, perquè els lectors acostumen a mirar-se només els titulars i les notícies alt per alt.

Quant a les estadístiques, és cert que els periodistes, jo m'hi incloc, els tenim un cert respecte, i per això a l'*e-notícies* no se n'ha publicat mai cap.

Per fer intel·ligible aquest llenguatge estadístic cal que els experts ens l'expliquin i no únicament amb dades, perquè és la manera que tenim d'entendre-ho i de transmetre-ho als lectors. N'és un exemple un article de Francesc Puigpelat, en què explica el dèficit fiscal català mitjançant una comparació entre Catalunya i una altra comunitat autònoma.

Dos altres exemples són notícies publicades a *La Vanguardia* i *El País*, com ara "Cataluña reducirá en 2009 el gasto per cápita en sanidad" o "Los gigantes de la economía mundial". Aquestes notícies són estadístiques parafrasejades, però malgrat això, la resta d'estadístiques estan condemnades a les pàgines d'economia. Això en la premsa en paper. En els diaris digitals, encara es fa més difícil adaptar les dades estadístiques perquè tinguin una bona acceptació per part dels lectors.

En resum, cal que les dades estadístiques que ens arriben siguin intel·ligibles, senzilles, per als periodistes, de manera que es pugui transmetre la informació al lector de manera entenedora.

¹² Aquesta ponència és una transcripció de l'enregistrament que es va fer durant les Jornades, degudament autoritzada per l'autor.

Disponibilitat lingüística en països plurilingües¹³

Guillem Rico

Doctor en Ciència Política per la UAB

Enquestar la població en un espai plurilingüe no homogeni planteja molts reptes. La finalitat de les enquestes és que la mostra sigui representativa, és a dir, que se'n puguin extreure dades que reflecteixin fidelment la població i que siguin comparables en contextos diferents.

En les poblacions plurilingües, però, enquestar la població planteja dos problemes. D'una banda, que la població no entengui la llengua amb què se li fa l'enquesta o que no la parli provoca que no es pugui arribar a un percentatge important de la població. D'altra banda, hi ha un problema de comparabilitat; és a dir, si es decideix fer una traducció del qüestionari per poder arribar a aquesta població, hi ha la possibilitat que es cometin errors, que es perdin matisos i, per tant, que hi hagi una distorsió de les preguntes, que no serien exactament iguals.

Per solucionar aquest problema, les pràctiques habituals que realitzen les empreses que es dediquen a enquestar aquest tipus de poblacions són: excloure la població que no pot entendre la llengua en què es fa l'enquesta (en el cas que aquesta població representi un percentatge reduït); fer una traducció de l'enquesta (en cas que aquesta població representi un percentatge important) o traduir l'enquesta espontàniament, cosa que passa sovint en les llengües que són properes, com per exemple a Catalunya.

En el cas de les enquestes telefòniques, per exemple, el Centro de Investigaciones Sociológicas permet a les persones bilingües —haurien de ser-ho totes, tot i que sovint no és així— fer la traducció espontània dels qüestionaris, sense una traducció prèvia.

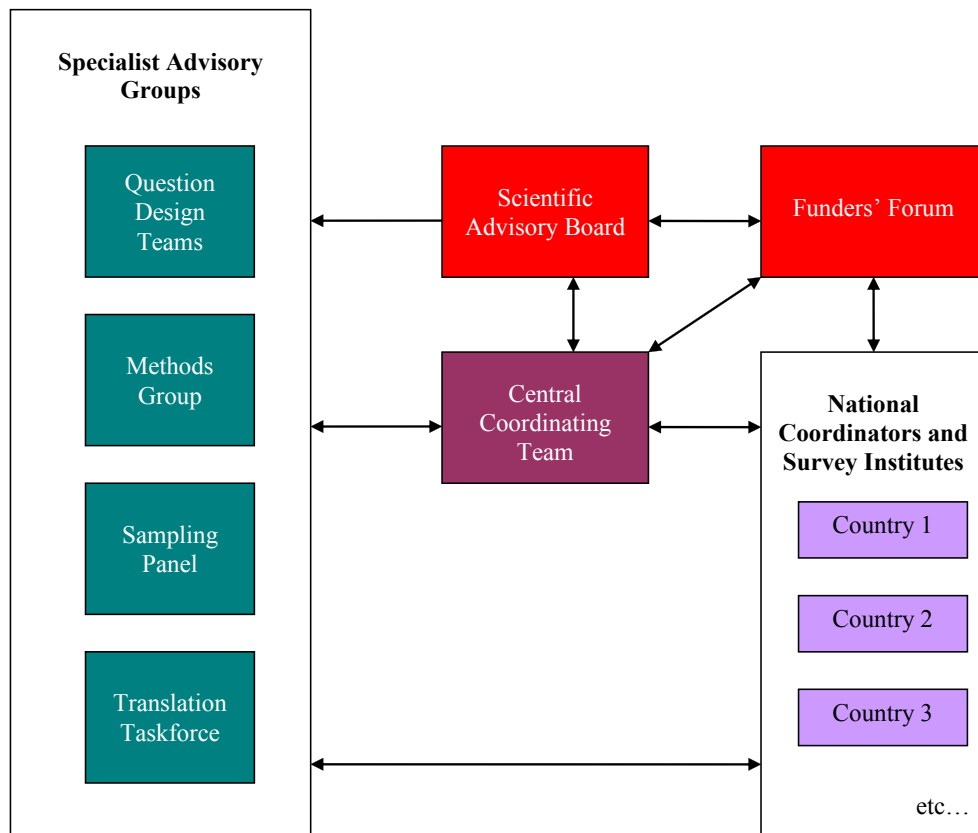
El problema de les traduccions, però, és que sovint no hi ha unes directrius establertes sobre com fer-les: sovint s'encarrega a un traductor i després es fa una revisió en diagonal. Sabem que les traduccions provoquen prou desviacions perquè s'hi hagi d'anar molt amb compte.

És per això que el projecte de l'Enquesta social europea ha proposat diverses solucions. Aquesta enquesta és un estudi comparat que es porta a terme cada dos anys en diversos països europeus i que recull dades sociodemogràfiques i actituds, valors, comportaments, de la població en tres àmbits: el social, el polític i l'econòmic. El valor d'aquesta enquesta és que té la voluntat de ser metodològicament molt rigorosa, en el sentit que pretén obtenir dades de països europeus que siguin comparables. Per això, aquesta enquesta recull mostres probabilístiques estrictament aleatòries. D'aquesta manera, s'eviten les quotes, per les quals no es pot calcular la probabilitat d'error de l'enquesta i, per tant, no hi ha problemes a l'hora de fer les inferències cap a la població. L'objectiu és que en la mostra seleccionada hi participi al màxim de gent possible, de manera que, si una persona no accepta fer l'enquesta, es comptabilitza com a perduda (no se substitueix), i passa a formar part del percentatge de no-resposta.

Quant a l'estructura organitzativa d'aquesta enquesta, funciona de la manera següent: la Unió

¹³ Aquesta ponència és una transcripció de l'enregistrament que es va fer durant les Jornades, degudament autoritzada per l'autor.

Europea finança la coordinació de l'estudi, però cada país n'ha de finançar l'execució. En el cas de l'Estat espanyol, la finança el Ministeri de Ciència i Educació, amb la col·laboració de la Generalitat.



Estructura organitzativa de l'Enquesta social europea
Font: www.europeansocialsurvey.org

Així doncs, les solucions que presenta l'Enquesta social europea als reptes ja esmentats són diverses. D'una banda, si una llengua és parlada com a llengua mare pel 5% de la població d'un país, cal traduir el qüestionari a aquesta llengua; això suposa que, a l'Estat espanyol, el qüestionari d'aquesta enquesta estarà redactat només en dues llengües: castellà i català.

D'altra banda, proposa un protocol de traducció de l'anglès (llengua original) a totes les altres llengües, anomenat *Ask the question*, que pretén l'equivalència exacta de continguts i no la traducció més literal. És a dir, que tant el receptor d'una llengua com el de l'altra han de percebre el mateix significat. Aquesta metodologia, doncs, s'allunya de l'anomenada *back translation*, que consisteix a traduir el qüestionari de la llengua de partida a una d'arribada, i d'aquesta a la de partida una

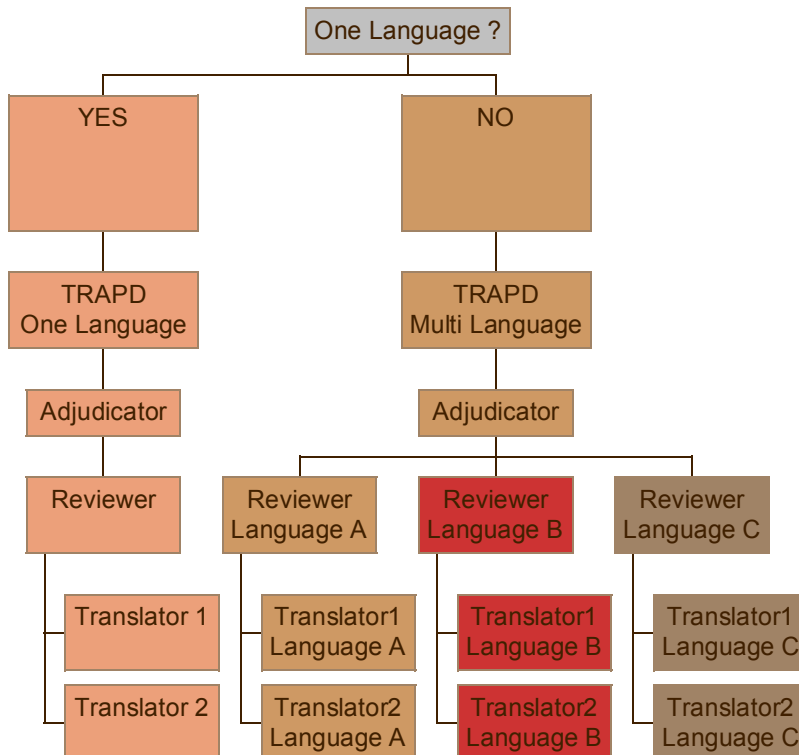
altra vegada, per contrastar diferències. El problema que té aquest tipus de traducció, però, és que sovint és massa literal i la llengua del qüestionari queda artificial. En l'Enquesta social europea, en canvi, es busca que el llenguatge de la llengua d'arribada sigui natural, i el biaix que pugui haver-hi es compensa per l'equivalència de significats.

Es tradueix a totes les llengües des de la llengua original i no de l'anglès cap a una i d'aquesta cap a una altra, de manera que s'evita que hi hagi més desviacions de les que hauria d'haver-hi. A partir d'aquí, se segueix un procediment anomenat TRAPD, que consta de diferents fases: traducció, revisió, adjudicació, pretest i documentació.

El procés de traducció no és individual sinó que el realitza un equip, format per traductors professionals, però també per altres persones que no ho són professionalment.

L'esquema següent ens mostra què passa segons si en un mateix país hi ha una llengua o n'hi ha dues.

Organisation of Translation for One Language or Several Languages

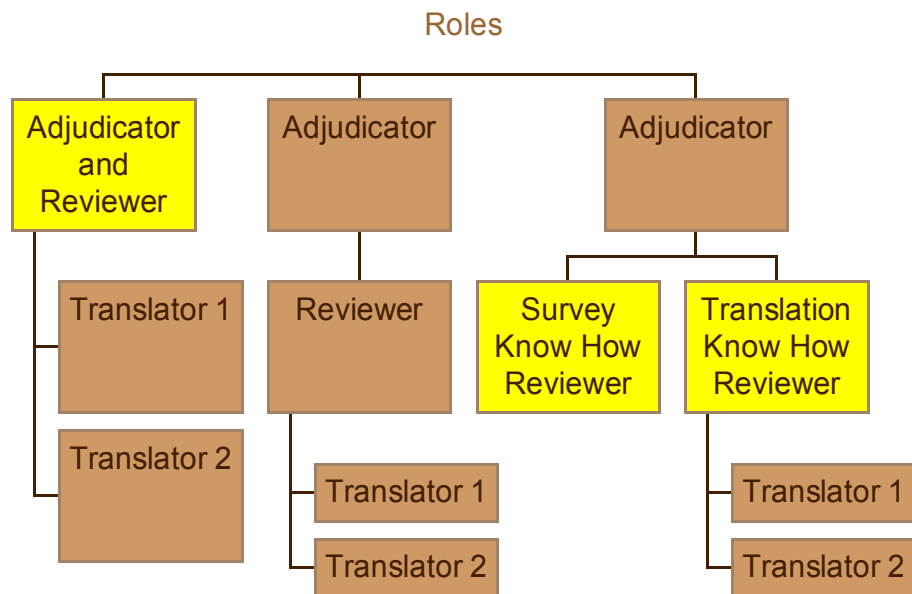


Organització de la traducció del qüestionari de l'Enquesta social europea
Font: Janet A. Harkness, "Round 4 ESS Translation Strategies and Procedures".

Tal com mostra l'esquema, per a cada llengua es realitzen dues traduccions independents i paral·leles. En el cas concret de l'Estat espanyol, per exemple, es porten a terme dues traduccions de l'anglès al castellà i dues de l'anglès al català.

Les traduccions es fan a més de per llengua, per país, de manera que a Bèlgica, per exemple, se'n fan dues, i a França també. Aquesta metodologia se segueix perquè es té en compte el context cultural, que és diferent en cada país.

De la traducció se n'encarreguen traductors professionals amb molta experiència, que tenen com a primera llengua la d'arribada, però que tenen també un coneixement molt bo de l'anglès. A aquests traductors se'ls donen unes indicacions de com cal traduir els qüestionaris. Les dues traduccions paral·leles es posen en comú amb un revisor que té un bon nivell d'anglès o de la llengua d'arribada i que, a més, és expert en disseny de qüestionaris. Es tria una traducció o s'agafa una part d'ambdues i després l'adjudicador (que de vegades és el mateix revisor, o el coordinador nacional de cada país) hi dóna el vistiplau.



Equip per a la traducció del qüestionari

Font: Janet A. Harkness, "Round 4 ESS Translation Strategies and Procedures".

Aquest procés comporta també una retroalimentació, perquè durant l'elaboració i revisió del qüestionari van sorgint preguntes i es van detectant ambigüitats que no s'havien previst a l'inici, cosa que afavoreix la millora de l'original.

A més, també es fa un pretest; és a dir, que es prova si el qüestionari original funciona: si és senzill d'entendre, etc. Habitualment aquesta prova es fa només amb l'original, però en l'Enquesta social europea es fa també amb les traduccions.

Finalment, es fa la documentació de tot el procés: les traduccions es passen a l'equip coordinador i s'enregistren els dubtes i les incidències que han aparegut al llarg del procés de traducció.

Els avantatges que comporta seguir aquest procediment de traducció són diversos. D'una banda, permet abraçar una mostra més gran. En el cas de les zones de parla catalana, per exemple, sembla clar que la majoria de la gent entendria el qüestionari en castellà, però probablement no estarien disposats a participar en l'enquesta si només fos en castellà. Per tant, com que interessa que la taxa de resposta sigui tan alta com es pugui, és millor traduir el qüestionari. D'altra banda, seguir un procediment d'aquest tipus permet millorar la qualitat de les traduccions i també del qüestionari original.

L'inconvenient principal que planteja aquest procés és el cost. Fer dues traduccions per cada qüestionari requereix una despesa molt més important que no pas si la traducció es fa espontàniament. Un altre inconvenient és que a vegades apareixen divergències entre les traduccions. Tot i això, s'ha comprovat que el sistema funciona i que les traduccions són equivalents quant a significat.

Exemple de divergències entre traduccions:

- Versió original anglesa

Now thinking about the European Union, some say European unification should go further. Others say it has already gone too far. Using this card, what number on the scale best describes your position?

- Traducció castellà

Pensando ahora en la Unión Europea, algunas personas dicen que la integración debería ir más lejos. Otras dicen que la integración ya ha ido demasiado lejos. Utilizando...

- Traducció català

Pensant ara en la Unió Europea, algunes persones diuen que la integració hauria d'anar més lluny. Unes altres diuen que la integració ja ha anat prou lluny. Utilitzant...

A part, hi ha altres inconvenients, com ara que de vegades hi ha petites desviacions que obliguen a coordinar les traduccions. Si s'han fet de manera independent, permeten comparar qüestionaris i corregir-los. A més, hi ha problemes quant a l'administració del qüestionari en català. Malgrat els esforços, a Catalunya molts pocs qüestionaris es van fer en català, i encara menys a les Illes Balears o al País Valencià.

D'aquí se'n desprèn que part del problema són les empreses a les quals s'encarreguen els qüestionaris, perquè el perfil d'enquestador va canviant, i actualment és molt difícil controlar que els enquestadors siguin bilingües.

Esforç institucional per un llenguatge accessible

Joan Abril Español

Assessor lingüístic de l'Idescat

1. Justificació
2. La revisió del model de llengua dels registres d'especialitat
3. El llenguatge estadístic i els seus diversos gèneres o productes
4. El paper de l'editor en la difusió dels productes estadístics
5. Criteris d'accessibilitat en terminologia, sintaxi o convencions
6. L'accessibilitat com a fil conductor del llibre d'estil de l'Idescat
7. El compromís de la institució en les bones pràctiques estadístiques i lingüístiques

Abstract

A l'hora de definir el model d'un llenguatge d'especialitat ens cal conèixer a fons aquest llenguatge i, per tant, disposar dels criteris gramaticals i terminològics que permetran classificar-lo a un nivell d'especialitat estàndard, és a dir, dins dels paràmetres de qualitat exigibles com a principi d'una bona pràctica estadística.

El llenguatge estadístic, així com el periodístic o el literari, també presenta diversos gèneres que es defineixen segons les necessitats comunicatives, és a dir, segons el format del producte (un informe, un document de treball, un tríptic, un llibre...) o el canal de difusió (paper, web, telèfon, presencial...).

També l'accessibilitat depèn en un grau molt elevat de l'establiment d'uns criteris rigorosos d'edició que permetin difondre els productes estadístics amb la qualitat necessària. En aquest sentit, cal que es doni una intersecció perfecta entre el projecte editorial, els productes i els destinataris. Per fer aquesta anàlisi de necessitats l'Idescat ha de disposar d'un Pla de publicacions, d'una política editorial i d'un llibre d'estil.

Un llenguatge és accessible si reuneix tots els mecanismes necessaris perquè sigui comprès fàcilment per un gran públic. En àmbits d'especialitat com l'estadístic, que és una ciència multidisciplinària, la terminologia s'ha de fixar d'acord amb criteris estàndards internacionals de comparabilitat i homogeneïtzació amb la terminologia d'altres llengües europees... En sintaxi o en convencions, cal cercar solucions sempre genuïnes, al marge del caràcter estrictament normatiu o no.

L'accessibilitat ha de ser el fil conductor no tan sols del llibre d'estil de l'Idescat, sinó del llibre d'estil de qualsevol institució que tingui com a principal objectiu la qualitat de la difusió estadística.

Finalment, l'Idescat, com qualsevol organisme estadístic, ha de tenir cura de la qualitat de les da-

des estadístiques que es difonen, dins la qual s'ha d'integrar la llengua mateixa, que ha de complir els criteris d'accessibilitat que s'han de prioritzar com a bona pràctica estadística i lingüística.

1. Justificació

El llenguatge de la matemàtica i de l'estadística presenta dades, resultats, metodologia, i les explica amb l'objectivitat que requereix la ciència. Ara bé, el llenguatge, com a varietat funcional d'una llengua, es defineix segons el canal de comunicació, l'especialitat terminològica, la intencionalitat i el nivell de formalitat. Aquests dos darrers factors, intencionalitat i nivell de formalitat són elements modulables per arribar a fer un llenguatge plenament accessible.

El propòsit de la meua ponència és precisament aquest: l'accessibilitat del llenguatge estadístic o, dit d'una altra manera, l'esforç que fa una institució com l'Institut d'Estadística de Catalunya perquè el llenguatge relatiu als resultats estadístics sigui al màxim d'accessible, és a dir, entenedor, clar i alhora precís, que no doni joc a cap mena d'ambigüitat i que alhora sigui tan permeable que un nen de 12 anys pugui entendre també l'IPC, el PIB o l'enquesta trimestral de cost laboral. Aquest "esforç" s'ha d'interpretar no pas com un suplici per al personal de l'Idescat, sinó com una oportunitat i alhora un gaudi que té com a objectiu un servei públic de qualitat.

Considero, però, que per arribar a aquest servei públic de qualitat, hem de ser autocrítics i alhora saber escoltar les opinions i observacions d'altres oficines estadístiques o professionals de diverses disciplines científiques. Per tant, us parlaré de la revisió del model de llengua dels llenguatges tècnics; del llenguatge estadístic i els seus gèneres (és a dir, diverses necessitats comunicatives segons tipus d'usuaris); també dels criteris d'accessibilitat en terminologia, però sobretot en sintaxi; del paper de l'editor en la difusió dels productes estadístics; de l'accessibilitat com a fil conductor del llibre d'estil de l'Idescat, i finalment del compromís de la institució en les bones pràctiques estadístiques i lingüístiques.

2. La revisió del model de llengua dels registres d'especialitat

La química, la geologia, el dret, la física quàntica, l'economia o l'estadística són llenguatges d'especialitat que tenen diversos punts en comú, un dels quals és la terminologia pròpia de la seva àrea d'especialitat, un llenguatge tècnic que ja s'ha tractat en la primera àrea científica. Els models de llengua dels registres d'especialitat s'han de revisar periòdicament, fins i tot la mateixa terminologia ha d'estar en un procés d'observació, revisió i proposta en funció de la llei de l'oferta i la demanda, és a dir, què m'ofereix la llengua, amb les regles de derivació i composició, i què ofereix la societat, d'acord amb un filtre de genuïtat lingüística. Per exemple, *palet* o *paleta*; *cistell de consum*; *complimentació* o *resposta*...

Comentem el primer exemple: el terme *paleta* normatiu contrasta amb el terme *palet* aparegut l'any 1983 a la 1ª edició del *Lèxic de la construcció* de l'Institut de Tecnologia de la Construcció de Catalunya (ITEC). En edicions posteriors, però, es va canviar el criteri, probablement per evitar una adaptació del manlleu poc adequada a les regles morfològiques entre el francès, la llengua exportadora del terme, i el català. En francès, el mot *palette* és femení (*la palette*), i, per tant, si hem de respectar el gènere femení, ha d'acabar en -a, és a dir *paleta*. Aquests criteris lògics d'adaptació d'un manlleu del francès, no casen, però, amb l'ús generalitzat de l'adaptació que en fan els es-

pecialistes de l'àrea de transport, que l'usen com a masculí a partir de la fonètica francesa, que és “*palet*”, la representació escrita de la qual, doncs, és *palet*, i en conseqüència el gènere del mot canvia automàticament a masculí: *el palet* en lloc de **la palet*. També el castellà, a través del *Diccionario del español actual*, recull el masculí (*el palet* i *el palé*). Finalment, l'ús entre els especialistes s'ha reflectit en la 2a edició del DIEC per mitjà d'una relació sinonímica entre *paleta* i *palet*, amb preferència pel terme masculí. Això, doncs, ens confirma la nostra tesi de la revisió constant de la ciència lingüística i estadística per tal de supervisar les regles o observar-ne l'excepcionalitat.

Però no només hem de tractar la terminologia d'una àrea d'especialitat quan parlem de revisar-ne el model de llengua. Hi ha altres elements que són tant o més importants, com ara la sintaxi o el mateix estil. Sempre que s'ha volgut definir l'estil d'un model lingüístic ha sigut com picar en ferro fred, perquè no se sap ben bé què és l'estil. Quan parlem de llibre d'estil, per exemple, crec que ens hauríem de decantar per la sàvia definició que va fer Teresa Cabré l'any 1989.¹⁴ Segons l'autora, “l'estil no és sinó el coneixement profund de la capacitat funcional de la llengua, el coneixement de la seva estructura i de les possibilitats que ofereix de ser manipulada (...). L'adequació a un determinat registre està estretament lligada a la utilització i a l'elecció dels recursos lingüístics i del vocabulari, però és difícil posar patrons massa esquemàtics (...). Aprofundint en la capacitat combinatòria de la llengua catalana, en definitiva, en la gramàtica, coneixerem els seus recursos estilístics.”

3. El llenguatge estadístic i els seus diversos gèneres o productes

Quan parlem de llenguatge estadístic, ¿hem de parlar només dels conceptes “bàsics” d'estadística?, ¿o ens hem d'estendre molt més i parlar d'economia, societat, llengua, models socials, ocupació, hoteleria, viatges, educació, cultura, esport, justícia..., tal com es desprèn del Pla estadístic de Catalunya 2006-2009?

A l'hora de tractar, però, el llenguatge de l'estadística o fins i tot de l'estadística aplicada o estadística oficial catalana, ens cal saber si els conceptes són prou transparents, ens hem de preguntar si el que els estadístics i lingüistes entenem ho entendreà de la mateixa manera l'usuari especialista o no especialista en estadístiques.

Ens cal, doncs, distingir la tipologia d'usuaris de les estadístiques, ens cal aplicar els criteris d'adequació dels diversos gèneres del llenguatge estadístic i, per tant, hem de fer una anàlisi de necessitats comunicatives d'acord amb el tipus de públic al qual ens adrecem.

No és el mateix editar una publicació d'estadística econòmica que un qüestionari, encara que sigui de la mateixa temàtica. El format de les preguntes, les diverses opcions de resposta, les diverses classificacions normatives.... han de complir uns requisits d'estandardització diferents d'una obra digital o en paper. Això ens porta a considerar el concepte bàsic d'adequació del registre o del gènere estadístic o col·lecció. I això ens porta també a reflexionar sobre la idoneïtat del producte segons el nostre públic objectiu, el que ha de rebre i entendre les dades.

14 Teresa Cabré (La formació lingüística de traductors i intèrprets, dins *Revista de Llengua i Dret*, núm. 13 (1989), pàg. 53 i 54).

4. El paper de l'editor en la difusió dels productes estadístics

També l'accessibilitat depèn en un grau molt elevat de l'establiment d'uns criteris rigorosos d'edició que permetin difondre els productes estadístics amb la qualitat necessària. En aquest sentit, cal que es doni una intersecció perfecta entre el projecte editorial, els productes i els destinataris.

Per fer aquesta anàlisi de necessitats l'Idescat ha de disposar d'un Pla de publicacions, d'una política editorial i d'un llibre d'estil, en el qual s'apleguen criteris generals i específics que ens permeten assolir un llenguatge estadístic que ha de complir els estàndards de qualitat.

Projectes editorials

Els projectes editorials dins l'Administració s'han d'inscriure necessàriament en un entorn que va més enllà de la publicació en paper, ha de ser un concepte de difusió que ha d'integrar la publicació multimèdia (paper, web, àudio, vídeo) i en tots els formats i entorns mediàtics (premsa, TV, ràdio, museus, cinema, publicitat...). L'estadística oficial ha d'arribar a la societat catalana amb uns nivells de difusió generosos perquè s'accedeixin a les dades estadístiques d'una manera clara, transparent i amb la capacitat de síntesi necessària.

Publicacions (paper i web)

El producte editorial estadístic s'ha de difondre amb la màxima claredat. Per tant, els productes estadístics s'han de definir d'acord amb els factors següents:

- gènere estadístic o col·lecció: documents de síntesi, estadística instrumental o assistencial, monografies...
- mitjà de difusió: web, paper (llibre, catàleg, tríptic...), mèdia...
- àmbit de difusió: Administració (sistema estadístic), empreses, universitats, ONGs, internacional...
- nivell de difusió: especialistes, escoles, MCM, destinataris en general...
- grau de formalitat del producte: tractament del contingut, del llenguatge, de la terminologia, és a dir, de l'estil.

L'Administració és servei públic i, per tant, s'ha de basar en criteris d'excel·lència. Això vol dir combinar els factors esmentats i considerar les raons que justifiquen la qualitat d'un producte:

- selecció acurada de títols
- professionalitat amb constant reciclatge dels tècnics
- contingut equilibrat de text, gràfics i taules
- disseny excel·lent
- encert en el programari (per al web i per a l'edició de publicacions)
- bona difusió o socialització de les dades estadístiques

L'editor, doncs, ha de vetllar perquè es donin els requisits imprescindibles, ha de vetllar, doncs, perquè el disseny d'una taula, la disposició gràfica, les notes metodològiques, els índexs, les preguntes dels qüestionaris, la distribució d'una publicació, etc. compleixin un objectiu bàsic, el més

important, que és que els diversos perfils de destinataris accedeixin als resultats estadístics, o a la presentació o explicació d'un treball de camp...

5. Criteris d'accessibilitat en terminologia, sintaxi o convencions

Un llenguatge és accessible en terminologia, morfologia, sintaxi o convencions quan aquests diversos plans de la llengua compleixen uns requisits de transparència.

En **terminologia**, a més de difondre el terme normalitzat, cal supervisar o replantejar termes que, malgrat el vistiplau del Consell Supervisor del TERMCAT, són objecte de debat entre els especialistes o entre els mateixos usuaris de l'àmbit en qüestió: *cistell de consum* vs. **cistell de la compra*, *gènere* vs. *sexe*, *carret d'anar a comprar* vs. **carret de la compra*, *tarja* vs. **targetó*...

En **morfologia o convencions**, un exemple d'accessibilitat el tenim en el plural de *CD-ROM*. Segons les recomanacions de l'opuscle editat per la Generalitat titulat *Abreviacions*, “les sigles de noms comuns no tenen flexió de gènere ni de nombre, però sí que varien els elements de la frase que hi concorden”, o també “el plural d'una sigla no s'especifica de cap manera, ni afegint una *s* al final ni duplicant-ne els components”. També s'hi diu que “tot i que el plural de les sigles no es marca gràficament, quan una sigla representa un plural, en general, la pronunciem en plural, tant si es llegeix amb la pronunciació sil·làbica com si es llegeix lletrejada o desplegada” (*ATS, CAP, DOGC, ONG, CD-ROM*...). Doncs bé, el criteri de l'Idescat és ajustar-se a l'ús oral. Si pronunciem “cederroms”, ¿per què hem de reprimir aquesta tendència natural a pluralitzar i a escriure *CD-ROMs*? El model de llengua de l'Administració no ha de ser pas conservadora, ha de ser còmoda, i genuïna alhora, naturalment, i ha de cercar la manera de combinar aquests criteris per tal de disposar d'unes solucions acostades als usuaris o “clients” de l'estadística.

En **sintaxi**, encara patim els embats injustificats de l'absència d'article amb el 2000 i anys següents. L'any 1999 l'Administració es va començar a preocupar pel criteri que havien d'emprar a l'hora d'expressar els anys 2000 i següents en la data d'un document, perquè resulta que fins ara podem dir “10 de febrer de o del 1999”, i ara només ens sona d'una manera natural “del 2000”, “del 2007”, en lloc “de 2000”, “de 2007”.

Aquesta “preocupació” neix de la voluntat de fixar un criteri uniformitzador amb el temporal 2000 i els anys següents. Des de l'Administració, però, els tècnics lingüistes són conscients que “de 2000” no sona gens bé, però en canvi insisteixen en el criteri de l'ús analògic amb les expressions com “de 1957, de 1729...”, sense arribar a tractar que l'ús natural amb article respon a criteris sintàctics. L'argumentació que dóna Política Lingüística és la coherència formal de les datacions de documents amb la localitat i la comparabilitat de dates que van amb l'any. Malgrat, doncs, reconèixer que “la forma sense article la trobem forçada”, es refermen en la proposta sense article, amb la qual cosa se sacrifica la naturalitat i espontaneïtat de l'ús dels parlants. Des de l'Idescat entenem, per raons d'accessibilitat, que no podem sacrificar la naturalitat de l'article amb l'any, i encara hi ha una altra raó de pes, i és que, tal com diu el Dr. Joan Solà (membre de la Secció Filològica de l'IEC), l'única solució viable o correcta és amb l'article.

En **sintaxi** optem per un ús flexible de l'ordre dels elements en la frase. Per tant, acceptem tant les tematitzacions –quan pel context són elements ja coneguts– del tipus “Al teatre(,) hi ha(n) anat 29,9% de la població de més de 15 anys” com “Un 29,9% de la població de més de 15 anys ha(n) anat al teatre”, que és l'ordre neutre de l'oració. En la qüestió dels relatius, defugim els possessius massa allunyats del referent. Així, a l'EPA un títol de taula com ara “Habitatges familiars la persona principal dels quals és ocupada” el canviarem per “Habitatges familiars on (o bé “en què”) la persona principal és ocupada”. El criteri general és evitar estructures complexes que alenteixin la lectura, encara que sigui per centèsimes de segon.

Una altra qüestió **convencional** i també **sintàctica** és l'ús del masculí com a genèric per representar els dos sexes. Des de l'Institut Català de les Dones, a través d'una publicació, s'ha fet campanya per visibilitzar el femení perquè es considera que el masculí genèric discrimina la dona. L'Idescat, com a organisme de l'Administració i per tant de servei públic, està plenament d'acord que calen mesures per evitar la discriminació de la dona en molts camps o àmbits de la societat, però això no vol dir que hagi de compartir a cegues criteris que poden emmascarar l'espontaneïtat de l'ús dels parlants i que podrien fins i tot fer perillar el nivell d'accessibilitat estàndard del llenguatge estadístic.

La *Gramàtica de la llengua catalana* de l'Institut d'Estudis Catalans (IEC) diu que “En els noms en què el gènere estableix oposicions de sexe, el masculí és la forma no marcada semànticament. El caràcter no marcat del masculí es pot constatar fàcilment en el valor extensiu d'aquest gènere en contextos plurals”. Parlem, doncs, de *matemàtics* o *estadístics* i, sense espai per a la discriminació, abracem efusivament tant *matemàtics* com *matemàtiques*, així com *estadístiques* i *estadístics*.

És a dir, segons la gramàtica normativa de l'IEC, el gènere masculí és el gènere no marcat i, per tant, des del punt de vista normatiu, els textos procedents tant de l'Administració pública com de les empreses o els particulars s'adeqüen plenament a l'obra prescriptiva de la Secció Filològica de l'IEC.

El TERMCAT, Centre de Terminologia per a la Llengua Catalana, el Diccionari de l'IEC, entre altres, utilitzen el gènere no marcat, així com la premsa escrita, la ràdio o la televisió, per raons d'economia del llenguatge. Vegem alguns exemples de definició d'un diccionari terminològic elaborat pel TERMCAT:

- **clau secreta** *f* Clau reconeguda per **un emissor** i **pels receptors** corresponents, que permet al primer de xifrar un missatge i als segons, de desxifrar-lo.
- **ciberespai** *m* Espai virtual en què **els internautes** interaccionen per mitjà d'ordinadors connectats a una xarxa telemàtica, generalment Internet.

Contràriament, però, en impresos i documents de redacció tancada, siguin de l'Administració o no, cal que s'hi apliqui el criteri genèric de la doble forma, o una fórmula equivalent, per als noms dels càrrecs o de noms anàlegs per tal d'evitar un ús discriminatori: *El secretari/ària, El director/a...* Una altra cosa molt diferent és estendre aquests usos en altres contextos, en què el masculí (singular o plural) és el gènere que es fa servir amb un valor plenament genèric: *Els farmacèutics de Barcelona fan vaga. Els grans són la veu de l'experiència. Els funcionaris de presons han actuat correctament.* Això, és clar, no impedeix de pensar en altres fórmules com ara *Les farmàcies de Barcelona fan vaga. Estadística econòmica dels establiments d'atenció social a la gent gran o El funcionariat penitenciari ha actuat correctament.* En qualsevol cas, però, han de ser només propostes optatives, en lloc de prescripcions úniques per a l'Administració de la Generalitat.

- La publicació del Decret 137/2008, de 8 de juliol, pel qual s'aprova la Classificació catalana d'activitats econòmiques 2009 (CCAIE-2009) és una mostra que els usos masculins amb valor genèric, com ara *malalts mentals* i *drogodependents*, no vulneren pas el Decret 162/2002, de 28 de maig (sobre l'ús d'un llenguatge simplificat i no discriminatori), ni fa servir un ús androcèntric o sexista del llenguatge.

6. L'accessibilitat com a fil conductor del llibre d'estil de l'Idescat

El llibre d'estil d'un organisme com és l'Idescat no es limita a definir només uns principis generals de la seva política lingüística (criteris d'actuació lingüística en les operacions estadístiques, ús de les llengües en el web...), la redacció de quatre normes gramaticals o terminològiques, l'elaboració

d'un diccionari estadístic, sinó que va més enllà d'aquests projectes més o menys immediats. El llibre d'estil de l'Idescat és aplegar principis, criteris lingüístics, criteris d'edició i sobretot fer constar tots aquests criteris en suport digital o en paper per fer-ne la màxima difusió a tot el Sistema estadístic català.

La raó de fer constar per escrit tots aquests criteris, i fer-ne tanta difusió com es pugui, és senzilla i elemental: no haver d'"inventar" o "reinventar" criteris, poder actuar d'una manera coherent en col·leccions estadístiques en què ja s'han fixat unes convencions. Així com en terminologia estadística es parla sempre de "llar" i no de "casa" o "família", que serien altres conceptes, en el capítol de criteris d'edició la fe d'errates s'ha de consignar a la pàgina de crèdits de la versió PDF. Un exemple més de promoció d'un llenguatge accessible.

El llibre d'estil també ha de comprendre aspectes que poden ser aparentment col·laterals i que, en canvi, són importants per a un estil de qualitat o excel·lència d'un organisme, com ara l'extensió en la pregunta d'un qüestionari: cal mesurar la llargada si volem que l'enquestat interpreti de manera ràpida i correcta el que se li demana, com també la disposició de les taules tant en el contingut com en la presentació formal.

Un llibre d'estil ha de ser tot això i més, han de ser criteris —les normes les deixarem a banda— per convenció i convicció no tan sols del sistema estadístic, sinó també del públic en general, i, per tant, han de ser prou flexibles perquè gaudeixin de la complicitat de tots els destinataris.

A l'hora de pensar en el contingut d'un llibre d'estil com el de l'Idescat, cal que sigui molt obert, és a dir, que sigui revisat periòdicament d'acord amb el ritme de la societat de les TIC, de l'Administració electrònica, dels webs o les webs en col·laboració, etc. Per tant, hi haurà capítols que hauran de ser constantment revisats, corregits i augmentats, i uns altres que podran ser una mica atemporals.

Aquesta accessibilitat en el llenguatge estadístic es tradueix, doncs, en l'establiment d'uns criteris sobre convencions gràfiques i uns altres de més importants com són la terminologia, la sintaxi i la redacció.

D'una banda, les convencions gràfiques serveixen per acordar a l'Idescat i, de retop, dins el Sistema estadístic de Catalunya, unes solucions que siguin tan còmodes com sigui possible per afavorir l'accessibilitat. Així, tindrem els capítols següents:

- **Abreviacions:** abreviatures, sigles i acrònims.
- **Citació de fonts estadístiques:** capítol que recull la casuística per citar les fonts i la manera més àgil tant per a les de referència pròpia com aliena.
- **Majúscules i minúscules:** apartat que pretén simplificar la macrodiversitat d'aquesta convenció.
- **Referències bibliogràfiques:** criteris de citació documents i publicacions estadístics.
- **Criteris d'edició:** capítol important a l'hora de definir els diversos circuits de treball en l'edició de totes les publicacions de l'Idescat, periòdiques i no periòdiques, i on entren les diverses fases de correcció, maquetació, fe d'errates, ISSN, ISBN, dipòsit legal, codi de barres, política de preus, la distribució, etc. En aquest capítol hi tindria cabuda també un nou disseny de les taules més simple, així com els índexs, ara com ara massa atapeïts i que poden tenir un disseny amigable sense perdre contingut.
- **Web:** capítol molt important de la difusió estadística.
- **Política lingüística de l'Idescat:** tracta tant els criteris d'ús de les llengües al web com l'ús a l'hora de fer una operació estadística (enquestadors vs. enquestats, criteris d'ús de les llengües en els qüestionaris, els qüestionaris per la Xarxa, les llengües en la signatura dels convenis i cartes, etc.).

- **Criteris de redacció i estil:** s'hi inclou el recordatori o la revisió de la normativa gramatical. Un capítol rellevant perquè tracta qüestions de sintaxi o lèxic, que són fonamentals a l'hora de fixar un estil de llengua realment accessible. Per tant, es tracta l'estructura de l'ordre gramatical, les expressions temporals (molt importants en estadística), l'absència o l'evidència de certa informació, com ara "Catalunya" o "estadística" quan en les publicacions de l'Idescat o en el seu web es parla d'estadístiques de Catalunya, etc.
- **Criteris de traducció:** hi tenen cabuda criteris en les dues llengües d'arribada, com són el castellà i l'anglès. Què es tradueix i què no. Per exemple —llevat que sigui la citació d'una font estadística—, traduïm els noms dels organismes o de les institucions acadèmiques per raons també d'accessibilitat, és a dir, de funcionalitat. El que volem realment és que els destinataris de parla anglesa o castellana arribin a entendre de qui o de què parlem.
- **Signes de puntuació:** dins aquest punt, tot i que també pertany a la sintaxi, tractem els signes d'interrogació i d'exclamació. És obvi, per exemple, que en un qüestionari hi ha una bateria de preguntes adreçades als enquestats. Tot i així, i en preguntes complexes, és a dir llargues, és lògic que optem per un criteri flexible, malgrat que poc sistemàtic, com és el de Fabra —en lloc de la proposta al meu entendre massa simplificada de la Secció Filològica de l'Institut.
- **Criteris terminològics:** aquest capítol fa una remissió general al TERMCAT i recull la terminologia de l'estadística oficial, bàsicament el diccionari que l'Idescat està elaborant i que hauria de veure la llum el 2009, coincidint amb la celebració dels vint anys de l'Idescat.

7. El compromís de la institució en les bones pràctiques estadístiques i lingüístiques

Un organisme s'ha de fer bons propòsits en relació amb les "bones" pràctiques estadístiques o lingüístiques revisant seriosament què no fa prou bé i amb quins recursos i mitjans disposa per corregir-ho, de tal manera que aquestes pràctiques estadístiques serveixin realment per complir l'objectiu general del Pla estadístic de Catalunya 2006-2009, que és "assegurar el subministrament de dades necessari per al coneixement de la realitat, i per a la presa de decisions i la seva avaluació ulterior pels agents socials, econòmics i polítics".

La llengua és un instrument imprescindible i una disciplina de col·laboració per obtenir dades estadístiques i, en aquest sentit, les pràctiques lingüístiques, perquè les puguem qualificar de "bones", han d'encaminar-se vers tots els criteris que he esmentat per a l'accessibilitat dels usuaris, amb el propòsit normatiu que acabem de citar. Aquests criteris són instruments d'un servei públic que ha de revertir en noves planificacions i en noves millores per a la societat. I en aquest sentit el llenguatge accessible és una bona pràctica estadística i lingüística, amb un 5% d'inspiració i un 95% d'expiració.

La llengua com a indicador i instrument d'una operació estadística

Elena Sintés Pascual

Coordinadora del Grup de Recerca de la Direcció d'Usos del Temps

Des que Durkheim va assenyalar que els fets socials són observables i que aquests es poden operativitzar en forma d'indicadors empírics, les ciències socials han sigut prolífiques en l'elaboració de manuals, criteris metodològics i revisions epistemològiques sobre la forma en què aquestes disciplines s'han d'aproximar a la realitat. La recerca social es regeix per un conjunt de regles metodològiques dissenyades en l'ampli espai comprès entre els postulats teòrics i els de caràcter estrictament empíric. Del conjunt se'n poden extreure tres idees rellevants en el marc de les Primeres Jornades de Llengua i Estadística.

La primera és que les ciències socials mesuren realitats percebudes. El coneixement de la realitat que obtenim mitjançant els instruments de recerca social és un coneixement aproximat d'aquesta realitat. I és així en tant que la sociologia, i també l'estadística, assumeixen l'existència d'errors de mesura.

En segon lloc, el fet de ser conscients de no tenir un "coneixement veritable" sinó "aproximat o rectificat", en paraules de Gaston Bachelard (Bourdieu, 2001: 20), no ens ha d'impedir d'usar i treballar en el perfeccionament dels instruments de mesura. Duta a l'extrem, la prudència metodològica, la cerca de la realitat veritable, ens impediria conèixer res més enllà d'allò que percebem de forma intuïtiva, com aquell malalt esmentat per Freud que netejava contínuament les seves ulleres sense posar-se-les mai. Precisament, la creixent freqüència amb què se celebren congressos, s'escriuen articles i manuals, i s'organitzen cursos d'especialització metodològica dona compte de l'interès actual envers la metodologia i la qualitat en la investigació social.

Ara bé, i aquesta és la tercera idea, són escasses les obres que centren l'atenció en la influència que la llengua i el llenguatge tenen en els procediments metodològics i els instruments de recerca. "Herència de les paraules, herència de les idees", segons Brunschvicg, el llenguatge conté en el seu vocabulari i la seva sintaxi les categories que articulen la societat. I aquestes són emprades, sovint de forma inconscient, en la recerca social.

Tal com el defineixen Peter L. Berger i Thomas Luckman (1988), el llenguatge és una eina de significació cultural i un element de construcció de la realitat. Les prenocións, les categories i els seus significats són presents en tot procés de recerca, des de la construcció de l'objecte d'investigació fins a l'anàlisi dels resultats. Si bé hem d'acceptar que treballem amb instruments mancats de neutralitat, tampoc no ho podem d'ignorar i és convenient fer un ús crític i conscient del llenguatge, de les paraules i la sintaxi, de les categories amb què construïm els instruments de mesura.

El llenguatge en les operacions estadístiques: de la realitat percebuda al coneixement aproximat

L'objecte d'aquesta ponència és la revisió del paper de la llengua en les operacions estadístiques.

Comencem per caracteritzar què és una operació estadística, de quines parts es compon i en quines d'aquestes la llengua hi té una influència rellevant. Una operació estadística es pot definir com el conjunt de processos necessaris per a l'obtenció de resultats estadístics sobre una temàtica determinada a partir de l'obtenció d'un conjunt de dades individuals. Les operacions estadístiques es desenvolupen en dues grans fases: una primera, de plantejament, disseny i programació de la investigació, i una segona, de desenvolupament i aplicació.

Les operacions estadístiques segueixen, en termes generals, cinc grans etapes:

1. Plantejament de l'objecte. En aquesta etapa es determina l'objecte d'estudi, s'estableixen els objectius de la recerca i es formulen les hipòtesis de treball.
2. Disseny de la investigació. Es tracta de l'etapa d'establiment de la metodologia més adient per assolir els objectius i de definició de les característiques tècniques de la investigació. En les operacions estadístiques, on s'han d'obtenir dades individuals, es dissenya el protocol d'obtenció de dades, juntament amb tots els materials necessaris per a la recollida de la informació. En cas de tractar-se d'una enquesta, també és el moment en què es dissenya la mostra.
3. Obtenció de les dades. És el període de treball de camp, en el qual, mitjançant el protocol de recollida de la informació, s'obtenen les dades individualitzades i s'enregistren en una base de dades, després dels processos de codificació i gravació de les dades individuals.
4. Tractament i anàlisi de les dades. Es tracta de l'etapa d'aplicació dels procediments estadístics per tal d'obtenir i interpretar els resultats de la investigació.
5. Presentació dels resultats. Finalment, per bé que podria formar part de la fase anterior, els resultats es preparen per ser difosos.

Les operacions estadístiques, en tant que instruments analítics, estan subjectes a les categories de percepció dels investigadors i de la societat. La recerca social aborda "realitats que ja han estat anomenades, classificades, realitats que tenen noms propis i noms comuns, títols, signes, sigles" (Bourdieu, Chamboredon i Passeron, 2001: 65). El plantejament de l'objecte d'estudi no s'escapa d'aquest condicionant, com tampoc no són neutrals ni el disseny dels instruments de recerca ni la interpretació dels resultats.

Sovint, les operacions estadístiques recorren a la tècnica d'enquesta per obtenir la informació. L'ús de qüestionari com a instrument bàsic de captació de la informació té un interès rellevant en el context d'aquesta Jornada, ja que parteix de la interlocució entre persones i, per tant, de la construcció lingüística. Amb la presència d'un enquestador o sense, la interlocució es dona en tots els tipus d'enquesta (entrevista telefònica, presencial, autoemplenada), ja que el qüestionari és, en realitat, el conducte que uneix la persona que investiga amb l'enquestada.

És evident, doncs, que l'elaboració del qüestionari és central en la investigació social i en moltes operacions estadístiques. Des del plantejament de les hipòtesis i la formulació dels objectius d'investigació fins a l'especificació de les preguntes i les variables que se'n deriven, el llenguatge esdevé l'instrument material fonamental. Si més no, es pot afirmar que una acurada tria i formulació de les preguntes i respostes, entre d'altres aspectes, és una condició *sine qua non* per obtenir una informació estadística vàlida i fiable. Tal com adverteix Díaz de Rada, del qüestionari en depèn "en primer lloc, que s'obtingui la informació que es desitja i no una altra, i, en segon lloc, la no-influència en la resposta dels entrevistats" (Díaz de Rada, 2005: 19).

La literatura metodològica posa l'èmfasi en el fet que l'entrevista mitjançant qüestionari és un mètode estandarditzat, és a dir, té com a característica principal que es fan les mateixes preguntes a tots els entrevistats, en el mateix ordre i en una situació social similar (Díaz de Rada, 2005: 18). L'afirmació parteix de la creença que si cada entrevistat s'exposa a un mateix qüestionari, les

diferències en les respostes s'han de poder interpretar com un reflex de les diferències socials (Fowler i Mangione, 1990: 14). Amb tot, el supòsit d'“igualtat de condicions” no es garanteix sempre. Un dels biaixos més freqüents són les diferències en la comprensió de les preguntes i dels conceptes en funció de l'adscripció social de les persones entrevistades (entre molts d'altres factors, els significats de les paraules poden variar segons l'origen geogràfic i cultural, el nivell educatiu o l'edat dels individus).

La qüestió, doncs, és doble. En primer lloc, com podem garantir que les preguntes són interpretades i compreses de la mateixa manera per tots els entrevistats? Aquest assumpte està relacionat amb el context de significat i d'interpretació en què es realitza l'enquesta i en l'heterogeneïtat o homogeneïtat del significat de les paraules i de les preguntes, tal com veurem més endavant.

I en segon, com podem minimitzar els errors assumits en el procés de disseny i desenvolupament de les operacions estadístiques? Aquí, la cura en el llenguatge utilitzat, la construcció de les preguntes i les respostes, la revisió dels conceptes emprats i dels seus significats n'és la clau.

L'ús del llenguatge en els qüestionaris

El disseny d'un qüestionari i dels seus continguts s'estableix en funció de l'objecte d'anàlisi. A partir dels objectius definits a priori, es genera una llista de temes a tractar en el qüestionari. De cada tema se'n deriven una sèrie de variables, preguntes i categories. Vegem-ho per parts. En el qüestionari es recull informació referida a una sèrie de temes, dels quals s'han escollit un nombre determinat de variables sobre les quals es vol obtenir informació. Les preguntes són una expressió manifesta, normalment en forma d'interrogació, mitjançant les quals es recull la informació. Es pot donar el cas que d'una pregunta se'n derivi una variable, però també són possibles altres tipus de situacions com, per exemple, que hi hagi variables que no requereixen cap pregunta (les variables d'observació en són un cas, com el municipi de residència), o que d'una pregunta se'n derivin diverses variables. Pel que fa a les categories, són les diferents respostes associades a una variable determinada.

Les preguntes constitueixen el principal element del qüestionari. La fiabilitat de la recerca depèn en bona mesura de l'èxit en el procés de recollida de les dades i, per tant, d'un bon enunciat de les preguntes. És important assegurar-se que les preguntes es refereixen amb exactitud als temes sobre els quals es vol informació i que el tipus de pregunta sigui el més adequat per tal d'obtenir l'indicador desitjat. S'ha de comprovar que les preguntes estiguin correctament formulades, amb un estil de llengua adequat per als entrevistats, que no incloguin biaixos i que l'ordre amb què es formulen en facilita la comprensió. La realització d'un test del qüestionari és útil per assegurar que la redacció és correcta. Finalment, conèixer les característiques de la població objecte d'estudi (si és més aviat jove o gran, d'un estatus socioeconòmic determinat, el seu origen geogràfic, etc.) és important a l'hora de prendre decisions lingüístiques.

La redacció de les preguntes és un dels moments més decisius i que sovint ocupa més temps en les fases preparatòries de les operacions estadístiques. La forma de redacció de les preguntes depèn de l'àrea temàtica a investigar, el perfil de la població entrevistada, el tractament estadístic que se'n vol fer i el tipus d'indicador que es vol obtenir, així com el tipus de procediment d'enquestació. De fet, en algunes investigacions s'ha provat que petits canvis en la redacció de les preguntes poden ocasionar grans diferències en les respostes (Blaxter, Hughes i Tight, 2000).

Sudman i Bradburn (1974) assenyalen el disseny del qüestionari com la principal font d'error de les estadístiques. Es poden distingir diversos tipus d'errors originats per la formulació de les preguntes i respostes, alguns dels quals tenen a veure amb la redacció de l'enunciat, sobretot els “errors d'especificació”, deguts a una desigual interpretació de la pregunta per part de les persones

entrevistades. I també dels enquestadors, quan n'hi ha: els biaixos poden derivar d'una mala interpretació dels enquestadors i enquestadores si, a priori, no se'ls han explicat prou bé els criteris, conceptes i significats. És convenient redactar un manual per homogeneïtzar criteris, amb totes les anotacions i aclariments necessaris.

Abans, però, d'entrar en la formulació dels enunciats de les preguntes, un altre aspecte a tenir en compte és el tipus de preguntes existents segons la classe de resposta que es vol obtenir. Seguint la tipologia establerta per Azofra, les preguntes poden ser (Azofra, 2000: 10):

- Preguntes obertes-tancades-semiobertes
- Preguntes dicotòmiques, categoritzades o d'escala ordinal, d'escala numèrica, de valoració
- Preguntes directes-indirectes
- Preguntes de resposta única o múltiple

El criteri més genèric és que les preguntes han d'estar formulades de mode que la resposta sigui directa i referida a la informació desitjada. Per això cal que la redacció sigui tan concreta i específica com sigui possible. Per exemple, la pregunta “Va votar vostè a les darreres eleccions?” té un problema de manca d'especificitat, ja que no concreta a quines eleccions es refereix i la persona entrevistada podria contestar referint-se a les eleccions municipals, autonòmiques, estatals o comunitàries.

En aquest mateix sentit, també és important especificar les referències temporals sempre que sigui necessari. Així l'enunciat “Quina és la durada exacta del seu contracte actual?” pot prestar-se a confusió. Amb aquesta formulació no se sap bé si es pregunta per la durada total de contracte —des de la data d'inici fins la data final—, o si es pregunta pel temps que fa que la persona entrevistada està contractada —des de la data d'inici fins a la data de realització de l'entrevista.

La precisió en el llenguatge és sobretot important en les preguntes que fan referència a períodes de temps. No és el mateix preguntar sobre un hàbit realitzat regularment, o en el dia anterior a l'enquesta, en l'última setmana, en l'últim mes o en l'últim any.

Una de les regles més generals però alhora més rellevants és la d'utilitzar un llenguatge senzill, que pugui ser comprès per totes les persones entrevistades, sense paraules tècniques, abreviatures ni sigles. També s'han d'evitar les paraules abstractes, com tipus, classe o espècie. Per exemple, una pregunta com “A quin tipus de col·legi porta els seus fills?”, sense respostes suggerides, pot entendre's com a titularitat (és públic, concertat o privat), grandària (és gran, mitjà o petit), o es pot referir a si és religiós o laic, etc. És millor formular la pregunta així: “Els seus fills i filles van un col·legi públic, concertat o privat?”.

La indefinició dels conceptes és un problema força freqüent. La següent pregunta genera molts dubtes: “Creu vostè que Catalunya hauria d'apropar-se als horaris europeus?”, amb possibilitat de respondre “a) Sí, b) No, c) NS/NC”. Què s'entén per “horari europeu”? N'existeix un de sol? I significa el mateix per a tothom?

Les preguntes han de ser clares i enteses per tots els enquestats en el mateix sentit. Cal tenir present que les paraules no tenen sempre el mateix significat per a tothom. Aquest és precisament el problema de la següent pregunta: “Amb quina freqüència va vostè al cinema?” on les respostes són: “a) Molt, b) Bastant, c) Poc, d) Molt poc, e) Mai”. En aquest cas les categories de resposta poden tenir significats diferents per a cada persona. Així per a unes persones “anar un molt al cinema” pot significar anar-hi un cop al mes, mentre que per a una altra pot ser anar-hi cada setmana. Per tal d'evitar aquesta confusió, és recomanable que l'investigador determini uns criteris comuns per a tothom i com més neutres, millor. En aquest exemple, les opcions de resposta haurien de recollir la quantitat de dies al mes.

El mateix tipus de problema es troba en l'enunciat "Utilitza vostè regularment el transport públic?", on la paraula "regularment" pot ser entesa de formes diferents (cada dia, alguns dies a la setmana, etc.) i el concepte de "transport públic" pot crear confusions si no s'acompanya d'alguna especificació (la persona entrevistada pot dubtar, per exemple, de si el taxi és o no un transport públic). De nou, seria millor preguntar "Quants dies a la setmana utilitza vostè el transport públic?", explicitant quins mitjans s'inclouen sota el concepte de transport públic.

El problema de l'heterogeneïtat de significats és especialment problemàtica en les investigacions internacionals on s'utilitza un mateix qüestionari en diferents països (Beltrán, 1997). En la traducció dels qüestionaris s'han de tenir en compte les dificultats que poden sorgir de la manca d'equivalència cultural i lingüística en la definició de marcs conceptuals i de categories d'anàlisi.

Malgrat es recomana que les preguntes siguin curtes i concises, en algunes ocasions és convenient contextualitzar-les amb frases aclaridores o amb la definició del concepte quan aquest no pot ser entès per tothom. La contextualització és una bona solució al problema de l'heterogeneïtat de significat, ja que situa als enquestats en un marc cognitiu únic i homogeni. Per exemple, en una qüestió sobre violència es pot fer una definició prèvia: "Ara parlarem sobre la violència a la parella o al matrimoni, és a dir, sobre les agressions físiques, sexuals o psicològiques que es produeixen entre persones que viuen o han viscut juntes, en parella: En termes generals, creu vostè que aquest tipus de violència és freqüent a Catalunya? Valori-ho en una escala de 0 (no n'hi ha gens ni mica) a 10 (n'hi ha moltíssima)".

S'han d'evitar les preguntes esbiaixades, que conviden implícitament a donar una resposta determinada. Per exemple, en la pregunta "La majoria dels catalans s'oposen a la guerra d'Irak, ¿comparteix vostè aquesta opinió?", és obvi que si a algú se li afirma que la majoria de població pensa d'una manera determinada, tendirà a contestar el mateix.

Les preguntes han de ser formulades de forma positiva, evitant els enunciats que expressen negacions. La pregunta "En quin dels següents llocs no s'hauria de permetre fumar?" s'hauria de substituir per "En quin dels següents llocs s'hauria de prohibir fumar?".

Una altra situació que també pot generar confusió de significats és l'ús de frases en sentit positiu i negatiu de forma intercalada. Per exemple, "Digui'ns si us plau la freqüència amb què el seu treball implica alguna de les següents situacions: a) Tenir espai suficient per treballar, b) Treballar en zones de molt difícil accés per a les mans, c) Il·luminació inadequada pel treball que realitza" i les possibilitats de resposta són "a) Sempre, b) Moltes vegades, c) De vegades, d) Només alguna vegada, e) Mai". En aquest cas, qui respon "sempre" en la primera frase està donant un sentit positiu "sempre tinc suficient espai per treballar", en canvi, la categoria de resposta "sempre" en les dues següents té un sentit negatiu "sempre treballo en zones de molt difícil accés per a les mans" i "sempre tinc una il·luminació inadequada".

Cada pregunta ha de contenir una sola idea. Un error molt comú és el d'incloure dues preguntes en una: "Està vostè d'acord que el govern hauria d'invertir més en infraestructures i menys en seguretat?" Algú pot considerar que s'ha de gastar més en infraestructures però no estar d'acord que s'hagi de reduir el pressupost de seguretat, o a l'inrevés. El més correcte seria formular dues preguntes separades.

Pot passar el mateix amb les categories de resposta. Per exemple: "Dels següents problemes indiqui el que considera més important" I entre les respostes hi ha "la situació laboral i l'accés a l'habitatge dels joves" o "l'exclusió social de la gent gran i de les persones discapacitades". En tots dos casos s'inclouen dues opcions diferents (situació laboral-accés a l'habitatge, gent gran-persones discapacitades). Una resposta afirmativa significa que s'està d'acord amb totes dues o només amb una? I el no, què vol dir? De nou, és molt millor separar els conceptes.

Pel que fa a les categories de resposta, s'han de complir dues condicions bàsiques: ser exhausti-

ves i excloents. Malgrat que aquestes són dues condicions profusament reiterades en els manuals metodològics, encara és habitual trobar qüestionaris que no ho aconsegueixen. Des de categories que no inclouen un ventall prou ampli de respostes, fins les que divideixen les respostes en intervals que se superposen, o les que contenen categories que s'inclouen unes en les altres. Un exemple és el següent: “Respecte al darrer diari que ha llegit, l'ha comprat, l'hi han regalat o prestat, o l'ha consultat per Internet?” amb les opcions de resposta “a) Compra, b) Gratuït, c) Internet”. En aquest cas el problema és que les respostes no són excloents, ja que la consulta per Internet pot ser amb pagament per la lectura d'un contingut o gratuïta.

La conceptualització i denominació dels indicadors estadístics

El traspàs de la informació obtinguda mitjançant el qüestionari a una base de dades per posteriorment fer-ne el tractament estadístic i l'elaboració i difusió dels resultats és una altra fase on l'ús lingüístic requereix molta atenció. La traducció del llenguatge amb què parlem dels conceptes inicials al llenguatge estadístic no és sempre directa, sinó que s'ha de tenir present el valor de cada paraula i, sobretot, cal fer una revisió crítica dels estereotips inclosos en la llengua.

Tal com s'ha dit al principi, l'objectiu final d'una operació és la construcció d'uns indicadors empírics que donin compte d'un fenomen concret. Lazarsfeld va desenvolupar en els anys seixanta el mètode d'apropament a la realitat social a través de l'ús d'indicadors socials. L'elaboració d'indicadors es basa en allò que Lazarsfeld va denominar “imagery”, una construcció de la realitat fonamentada en l'observació de fets socials regulars. Boudon i Lazarsfeld van diferenciar quatre fases d'elaboració dels indicadors (Boudon i Lazarsfeld, 1985: 36-40):

- Representació literària de l'objecte, en què l'investigador només té una percepció abstracta de la realitat.
- Especificació del concepte, on aquest és disseccionat en components, aspectes i dimensions.
- Elaboració d'indicadors per a les diferents dimensions a analitzar.
- Formació d'índexs o sistematització de les dades.

Un indicador es pot definir així com la mesura operativa de les dimensions d'un concepte. Cada concepte i les dimensions que el formen s'han d'operativitzar per tal de ser mesurables, és a dir, s'han de convertir en variables i, al final, en indicadors. Definit l'indicador, aquest s'ha d'etiquetar, ha d'anar vinculat a una descripció del seu contingut. I aquí és on el llenguatge torna a tenir un paper important.

En qualsevol operació estadística s'ha de decidir el model de llengua i la terminologia concreta amb què es volen descriure els indicadors estadístics, els títols i les notes metodològiques que acompanyen un resultat, taula o gràfic, i la redacció de l'anàlisi de les dades. L'estil de llengua s'ha de definir en funció del tipus d'usuari a qui s'adreça. Quan la informació va dirigida a la població en general s'ha de tenir cura a utilitzar un model de llengua més accessible evitant tecnicismes, però mantenint un ús correcte del vocabulari i la sintaxi.

A banda de definir un estil de llengua que faciliti l'apropament entre la informació estadística i els seus usuaris, qüestió analitzada a bastament en altres ponències d'aquesta Jornada, s'ha de tenir en compte l'evolució dels significats socials de les paraules i de les connotacions simbòliques que s'hi associen. Els usos lingüístics transmeten identitats, valors, esquemes de percepció, estereotips, costums, en fi, recullen les construccions socials. Evitar les prenocios amagades dins el llenguatge no és fàcil i cal un exercici crític, de reflexió i anàlisi, del llenguatge amb què construïm els indicadors estadístics.

Una de revisions crítiques del llenguatge dels indicadors estadístics més consolidada és la de la supressió de l'androcentrisme o sexisme en l'ús de la llengua. La crítica feminista al caràcter androcèntric de la recerca ha posat en qüestió moltes de les categories d'anàlisi de les ciències socials. Organismes internacionals i governamentals s'han plantejat la necessitat de trobar fórmules equitatives que identifiquin i representin homes i dones en pla d'igualtat. Fruit d'aquesta revisió és ja molt freqüent que les estadístiques utilitzin el gènere gramatical femení i masculí en les seves nomenclatures (professor/a; advocat/da; treballador/a; ocupat/ada, aturat/ada), incorporin l'ús de fórmules genèriques (infància, joventut) i que se substitueixin formes terminològiques d'herència androcèntrica (mestressa de casa) per altres de més neutrals (tasques de la llar).

Però, tal com s'ha anat insistint, la complexitat del llenguatge i dels seus significats va una mica més enllà. Seguint amb l'exemple de la terminologia relacionada amb els rols de gènere, la revisió crítica no s'ha limitat al vocabulari, sinó també a les categories conceptuals que s'analitzen. Un concepte objecte de debat és el del "treball". Així, les estadístiques sobre condicions de treball i els indicadors relacionats amb l'activitat laboral contenen categories rebutjades des del feminisme, com la distinció entre "treball" i "no treball", on el treball domèstic i familiar és invisibilitzat i agregat en la categoria de "no treball". Un cas similar és el de les enquestes de mobilitat, on es distingeix la "mobilitat obligada" i "mobilitat no obligada", en què els desplaçaments relacionats amb el treball domèstic i familiar es consideren "no obligats".

La conceptualització dels indicadors estadístics és un reflex de les convencions socials, també científiques, de cada període històric. La denominació i definició dels indicadors canvien amb l'aparició de noves categories socials. Així, si recentment la incorporació massiva de les dones al treball remunerat ha promogut un canvi en el llenguatge de les estadístiques laborals, unes dècades abans, tot just després de la depressió de 1929, la categoria *pobresa* va passar a ser substituïda per la categoria *atur*, com a reflex d'un canvi d'orientació de les polítiques socials.

Com a reflexió final, cal insistir en la responsabilitat de l'estadística en la descripció lingüística dels paisatges socials. Tal com s'ha intentat demostrar en aquesta ponència, un ús acurat de llengua és imprescindible per a l'obtenció de dades estadístiques de qualitat. Ara bé, també cal posar èmfasi en la necessitat que aquest ús sigui conscient dels significats de les paraules i de les categories que analitza, dels mons simbòlics que construeix i comparteix amb la societat.

Bibliografia

- Alvira, F. (2004). "La encuesta: una perspectiva general metodológica". *Cuadernos Metodológicos*. Madrid: Centro de Investigaciones Sociológicas.
- Austin, J. L. (1981). *Cómo hacer cosas con palabras*. Barcelona: Paidós.
- Azofra, M. J. (2000). *Cuestionarios. Cuadernos metodológicos*. Madrid: Centro de Investigaciones Sociológicas.
- Beltrán, M. (1997). "Sobre la traducción: un viaje de ida y vuelta". *Revista Internacional de Sociología*, vol. 17.
- Berger, P. L. i Luckmann, T. (1988). *La construcción social de la realidad*. Barcelona: Herder.
- Blaxter, L.; Hughes, C.; Tight, M. (2000). "Cómo se hace una investigación". *Colección Herramientas Universitarias*. Barcelona: Gedisa.
- Boudon, R.; Lazarsfeld, P. (1985). *Metodologías de las Ciencias Sociales*. Barcelona: Laie.

- Bourdieu, P. (2001). *¿Qué significa hablar? Economía de los intercambios lingüísticos*. Madrid: Akal.
- Bourdieu, P.; Chamboredon, J-C; Passeron, J-C (2001). *El oficio de sociólogo*. Madrid: Siglo XXI.
- Chevry, L. (1960). *El cuestionario*. Barcelona: Ariel.
- Díaz de Rada, V. (2005). "Manual de trabajo de campo en la encuesta". *Cuadernos Metodológicos*. Madrid: Centro de Investigaciones Sociológicas.
- Durkheim, Émile (1987). *Las reglas del método sociológico*. Madrid: Akal.
- Fishman, J. (1995). *Sociología del lenguaje*. Madrid: Cátedra.
- Fowler, F. J. Jr.; Mangione, T. W. (1990). *Standardized survey interviewing: minimizing interviewer-related error*. California: Sage.
- García Ferrando, M; Ibáñez, J.; Alvira, F. (1989). *El análisis de la realidad social. métodos y técnicas de investigación*. Madrid: Alianza.
- Sudman, S.; Bradburn, N. M. (1974). *Response effects on surveys: a review and synthesis*, Chicago: Aldine.

L'estadística i la llengua com a elements de cohesió social: l'anàlisi de l'estructura social catalana amb les dades del PaD

Cristina Sánchez Miret

Directora de Continguts de l'Enquesta panel de desigualtats socials a Catalunya a la Fundació Jaume Bofill

Introducció

Aquest article és un resum d'una part dels resultats de la investigació sobre estructura social realitzada amb les dades de la primera onada (anys 2001-2002) del Panel de Desigualtats Socials (PaD) a Catalunya.

El PaD Catalunya és una enquesta panel —es repeteix cada any l'enquesta a les mateixes persones—, que du a terme la Fundació Jaume Bofill amb una mostra representativa de la població de Catalunya de 16 o més anys i que abasta 1.991 famílies i més de 5.000 individus en el treball de camp de la primera onada.

L'anàlisi dels grups socials realitzat amb les dades del panel vol donar compte de quins grups de classe social coexisteixen en l'estructura social de Catalunya a partir de l'anàlisi de les condicions de vida, usos, opinions i comportaments socials de la població. Mesurant quins són i com són els grups socials, qui en forma part i què els defineix, podem explicar el tipus de recursos i béns socials que tenen uns i no altres, i també en què s'assemblen i els iguala, i que els fa diferents i els separa.

La metodologia emprada s'ha basat en l'aplicació de tècniques estadístiques multivariants mitjançant el programa de tractament de dades SPAD (Système Portable pour l'Analyse des Données) a partir d'un disseny de recerca en dues etapes. Una primera d'anàlisis de correspondències múltiples per àmbits temàtics (ingressos, educació, habitatge, llengua, cultura política...) per resumir i estructurar la informació, atès el gran volum de variables d'anàlisi de què disposa el PaD. I una segona, l'anàlisi de correspondències múltiples final, on els resultats dels àmbits s'han incorporat, al costat de les variables clàssiques de l'estructura social, com a variables sintètiques d'anàlisi.¹⁵

En aquest article s'exposen els resultats de la tipologia de llengua elaborada amb les dades del primer qüestionari del PaD (treball de camp de la primera onada, anys 2001-2002) i la seva relació amb cada un dels grups de classe social identificats, acompanyada també de la caracterització d'aquests respecte de l'origen geogràfic i el posicionament en l'eix nacionalista català o espanyol.

Disseny analític i tractament de les dades

El procediment a establir en una anàlisi clàssica de l'estructura social consistiria a interrogar la realitat a partir d'una estructura preconcebuda de la relació entre grups socials i classes, i assenyalar

¹⁵ Vegeu Sánchez Miret, Cristina; Quintana Garrido, Núria. "Les classes socials a Catalunya: desigualtats en les condicions de vida dels grups socials". A: Diversos autors. *Estructura social i desigualtats a Catalunya*. Volum 1 (2005). Fundació Jaume Bofill.

les característiques de cada una de les classes socials així predefinides. El plantejament metodològic que s'ha fet al PaD és força diferent.

La nostra prioritat no ha estat imposar teòricament¹⁶ una posició de classe a cada un dels grups socials definits per la seva ocupació, sinó a la mateixa vegada que establim les condicions de vida, els valors i les actituds de cada un dels grups socials, establir també la seva posició de classe. Dit d'una altra manera, no només volíem caracteritzar els grups socials, també volíem que la nostra anàlisi confirmés la relació existent entre els diferents grups socials i les classes.¹⁷ Per aconseguir-ho es planteja una metodologia basada en l'anàlisi multivariant.

Les tècniques d'anàlisi triades han estat l'anàlisi de correspondències múltiples, la partició jeràrquica i l'anàlisi de la variància multivariant: per a les dues primeres s'ha utilitzat el paquet estadístic SPAD (Système Portable pour l'Analyse des Données) i per a l'última s'ha utilitzat el paquet estadístic R.

Aquest procediment analític es va dissenyar per aprofundir en les relacions de dependència que s'estableixen entre les variables quantitatives que tractem, i en particular per poder explicar com els diferents valors o categories de les variables es relacionen els uns amb els altres per tal de traslladar l'estructura jeràrquica d'un conjunt de variables socials a l'espai dels individus; en el cas que aquesta estructura existeix en les dades.

Atesa la riquesa de la base de dades del PaD no es va plantejar directament una anàlisi de correspondències múltiples per analitzar els grups socials a partir de tota la informació que conté el qüestionari (484 variables) sobre els diferents aspectes de les condicions socials, actituds i valors de la població. Primer es van realitzar diverses anàlisis de correspondències múltiples, una de cada un dels aspectes tractats per l'enquesta, inclosa la llengua, per establir quin són els eixos de desigualtat en cada camp social a explorar. És a partir d'aquests eixos resultants convertits en una variable tipològica de l'anàlisi de cada àmbit que es realitza l'anàlisi de correspondències múltiples final en un procés continu d'estructuració de la informació, per tal de poder abastar tota la riquesa i complexitat que ens aporta la informació recollida.

Les classes socials a Catalunya

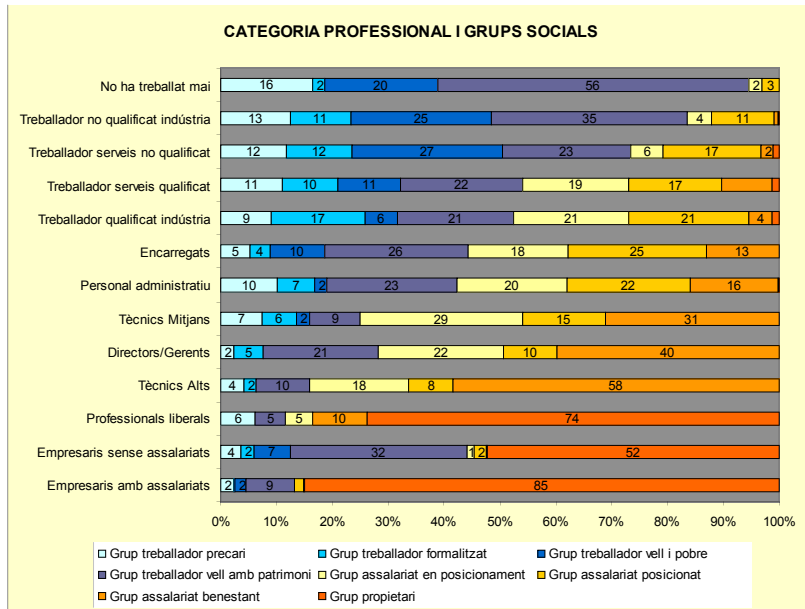
El resultat de l'anàlisi de les dades del PaD ens mostra vuit grups socials identificats i relacionats en una lògica de classes socials que plasma la persistència de l'estructura de classes i la seva importància en l'explicació de les desigualtats.¹⁸ Les tres grans classes identificades són: la classe mitjana propietària, la classe mitjana assalariada i la classe treballadora. La classe mitjana propietària està formada per aquells que treballen pel seu compte o formen part de l'empresa familiar, majoritàriament petits empresaris i autònoms de la indústria, els serveis i l'agricultura. La classe mitjana assalariada està formada principalment per tècnics alts i mitjans i una bona part del conjunt de persones que realitzen tasques administratives. I la classe treballadora bàsicament pels treballadors, qualificats i no qualificats de la indústria i dels serveis. Grups socials històricament coneguts, però que se'ns presenten caracteritzats amb nous elements i amb contrastos diferents entre ells i entre els mateixos grups que els formen respecte de les condicions de vida.¹⁹

16 Amb això no volem negar la importància del cos teòric existent sobre a quina classe social pertany cada grup social, sinó comprovar empíricament fins a quin punt aquesta relació es manté o ha canviat.

17 Vegeu anàlisis precedents amb aquesta mateixa metodologia: Sánchez 1994, Enquesta de la Regió Metropolitana.

18 Sumant-se als resultats d'altres investigacions d'estructura social que s'han produït en la mateixa línia. Veure Enquesta de la Regió Metropolitana de Barcelona 2000

19 La nostra anàlisi identifica vuit grups diferenciats, però que s'identifiquen en una lògica clàssica de classes socials. És a dir hi ha una clara divisió en les condicions de vida de tots aquells grups que són de classe treballadora respecte dels que ho són de classe mitjana. Alhora també queda plasmada la divisió interna dins de la classe mitjana, en la identificació de les



És important aclarir per què parlem de tres grans classes socials i no de vuit, si en l'anàlisi hem identificat tots aquests grups d'una manera diferenciada.²⁰ En primer lloc perquè els grups es relacionen a partir d'una lògica de classe, és a dir que hi ha una clara divisió en les condicions de vida de tots aquells grups que són de classe treballadora respecte dels que ho són de classe mitjana. La manera com s'uneixen i se separen els grups en el dendograma fa palesa aquesta divisió. També queda gràficament plasmada la divisió interna dins de la classe mitjana, en la identificació de les dues fraccions, l'assalariada i la propietària. Això no vol pas dir que l'explicació d'aquestes diferències internes dins de les classes no sigui important, sinó al contrari, explicarem en detall cada un d'aquests vuit grups perquè és el que ens permet acostar-nos millor a la riquesa de situacions i condicions d'existència del conjunt de la societat catalana.

En segon lloc, perquè hem comprovat amb una anàlisi de la variància multivariant la pertinença de considerar les tres classes identificades o de considerar una segmentació de l'estructura diferent. Els resultats de l'anàlisi de la variància multivariant²¹ reafirmen la nostra interpretació dels resultats de l'anàlisi de correspondències múltiples: la divisió de l'estructura social en tres grans classes, la classe mitjana propietària, la classe mitjana assalariada i la classe treballadora.²² Aquesta anàlisi ens permet mostrar que les distàncies entre els tres grups socials no són les mateixes, no es produeixen en la mateixa mesura i tampoc en els mateixos àmbits socials. Reforçant el que ja hem pogut copsar en el dendograma. D'una banda, les dues classes més properes són la mitja-

dues fraccions, l'assalariada i la propietària i en la identificació dels diferents estrats que la formen. Veure Sánchez Miret, Cristina i Quintana Garrido, Núria "Les classes socials a Catalunya: desigualtats en les condicions de vida dels grups socials". A: DA. Estructura social i desigualtats a Catalunya Volum 1 (2005) Fundació Jaume Bofill.

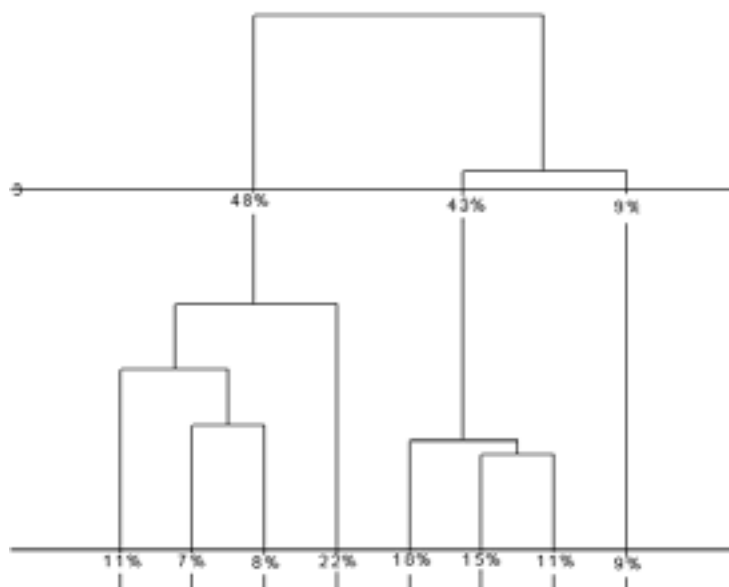
20 És important tenir present que malgrat que aquests vuit grups identificats en l'anàlisi són plenament significatius, hem de ser molt conscients que la segmentació dins de les classes no és que pugui arribar fins a l'infinit, però sí que és cert que pot ser més gran o més petita segons el criteri teòric i les representativitats estadístiques considerades.

21 L'anàlisi de la variància multivariant s'ha fet a partir d'un Tukey Contrasts i aquestes dues classes socials són diferents en 22 dels 42 eixos analitzats.

22 Hem realitzat tres anàlisis de la variància diferents. Un considerant per separat els vuit grups resultants de l'anàlisi de correspondències múltiple, un considerant només els tres grups que hem identificat com a classes i un de les posicions de treball mercantil dels individus. De tots aquests, ha estat el de tres grups, és a dir el de les classes socials, el que ha mostrat una significació més alta de la variabilitat entre els grups. És a dir, és a partir de les tres grans classes socials com millor podem explicar semblances i diferències en les condicions de vida de la població a Catalunya.

na assalariada i la mitjana propietària, que mostren menys diferències significatives²³ entre elles. D'altra banda, les diferències més grans les trobem entre la classe mitjana assalariada i la classe treballadora.²⁴

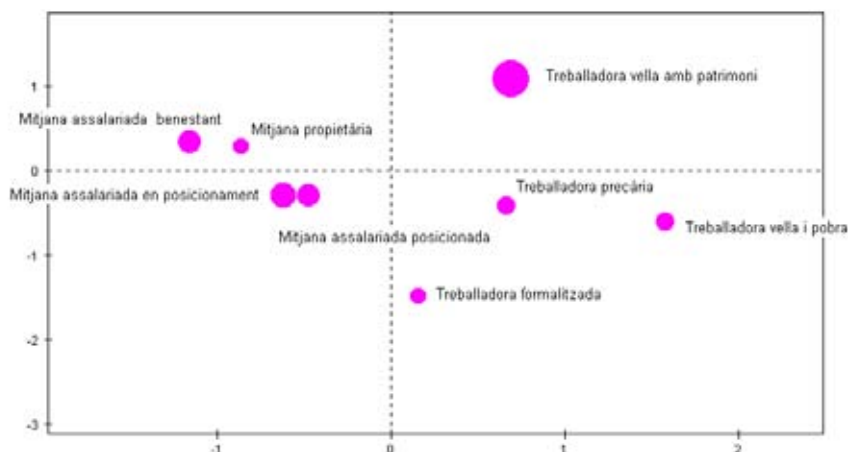
DENDOGRAMA



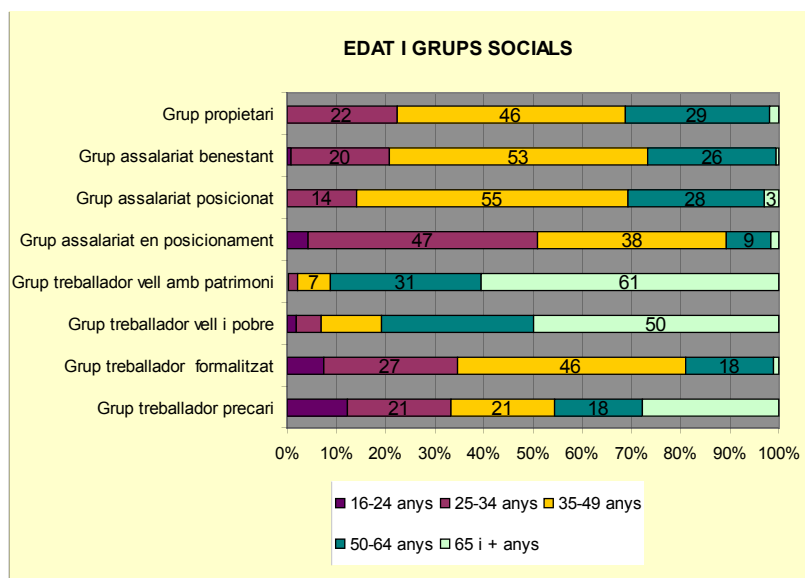
El dendograma il·lustra el pes i les distàncies (a partir de les semblances i diferències en les condicions, actituds i comportaments de les persones) de les tres classes socials. Semblances i diferències que explicarem més endavant respecte de la llengua. Ara cal que ens aturem en el tipus d'estructura que dibuixen els grups. El dendograma ens dona una aproximació gràfica de les relacions estructurals que s'estableixen entre elles –entre les tres classes– i entre els segments o grups interns que s'han identificat dins de cada classe social. Vegem-los en detall en el següent gràfic.

²³ Vegeu Sánchez i Quintana 2005.

²⁴ Aquestes dues classes es diferencien en 36 dels 42 eixos analitzats, és a dir pràcticament en tots. La classe mitjana propietària també mostra un alt grau de diferència amb la classe treballadora: es diferencien en 27 dels 42 eixos.

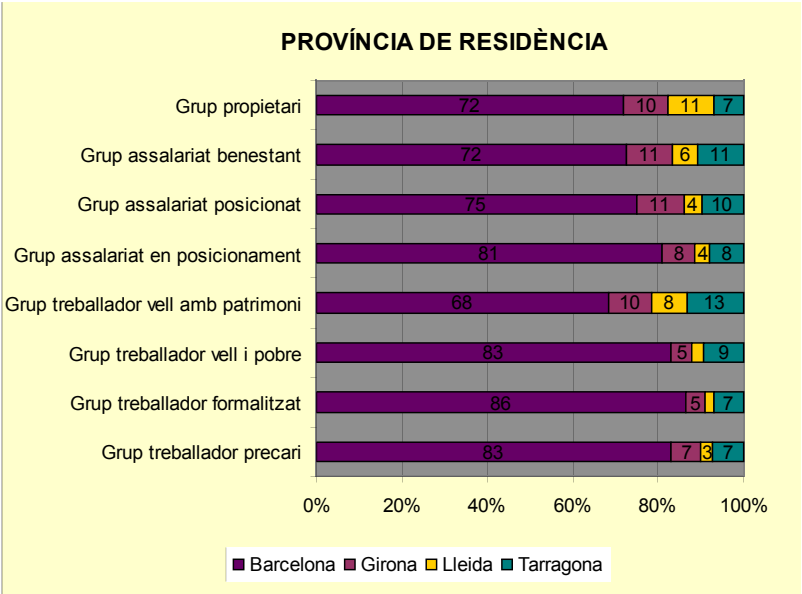


Els grups socials que componen les classes es defineixen bàsicament, a partir de les etapes del cicle vital dels membres que els componen, però d'una manera diferent en els grups de classe treballadora que en els grups de classe mitjana. A la classe treballadora hi ha una clara divisió entre els vells i aquells que no ho són, malgrat que en el grup que hem anomenat precari, comparteixen les mateixes condicions de vida uns i altres. Aquesta segmentació per edat no es reproduceix d'igual manera en els grups de la classe mitjana assalariada, malgrat que els grups que la formen també tenen mitjanes d'edat diferents.

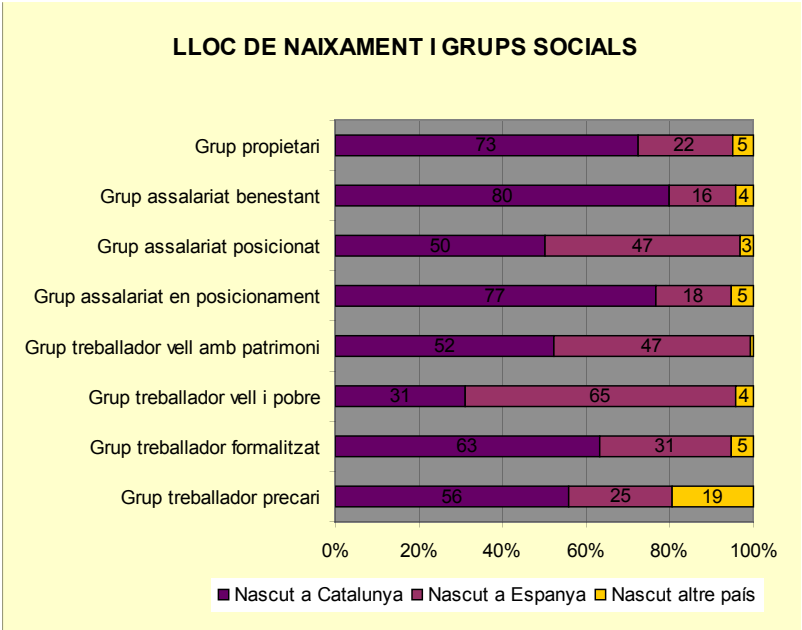


A la classe treballadora hi trobem quatre grups diferenciats pel cicle vital que marca les diferents etapes de relació amb el mercat laboral i amb la vida reproductiva de les persones. Dels quatre grups socials que la formen hi ha dos grups molt envellits, el grup vell i pobre (9,79%) i el grup amb

patrimoni (22%) amb la majoria dels seus membres jubilats, però amb diferent situació econòmica i lloc de residència.



Dels altres dos, hi ha un grup fortament marcat per l'etapa reproductiva dels seus integrants, el grup formalitzat (8,24%) i un altre grup, el grup precari (9,71%), en un clar procés de posicionament en el mercat laboral i en la societat, molts d'ells perquè no són nascuts en aquest país.

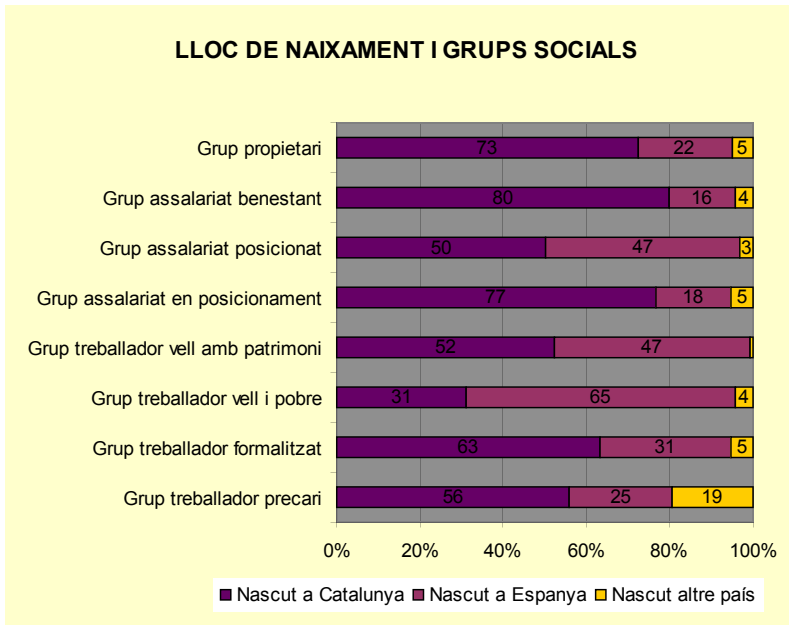


Els grups interns de la classe mitjana assalariada responen a una segmentació per etapes diferents de la trajectòria vital i per nivells de benestar, que donen com a resultat tres grups socials diferents: la classe mitjana assalariada en posicionament (14,53%), la classe mitjana assalariada posicionada (13,59%) i la classe mitjana assalariada benestant (12,95%).

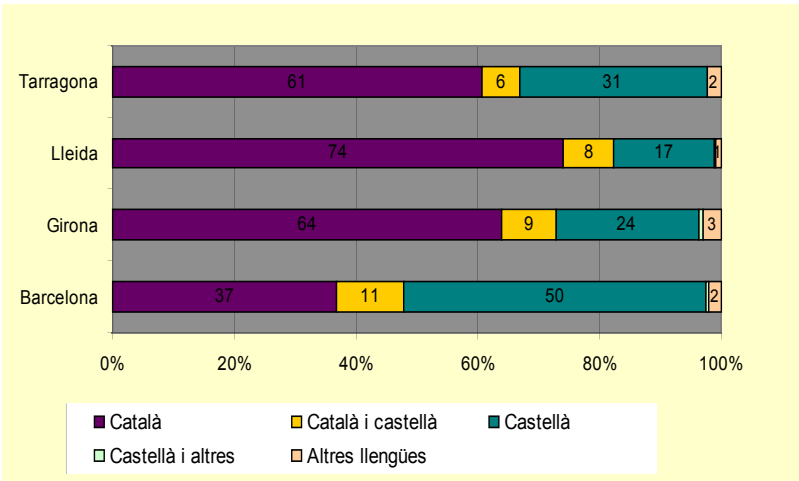
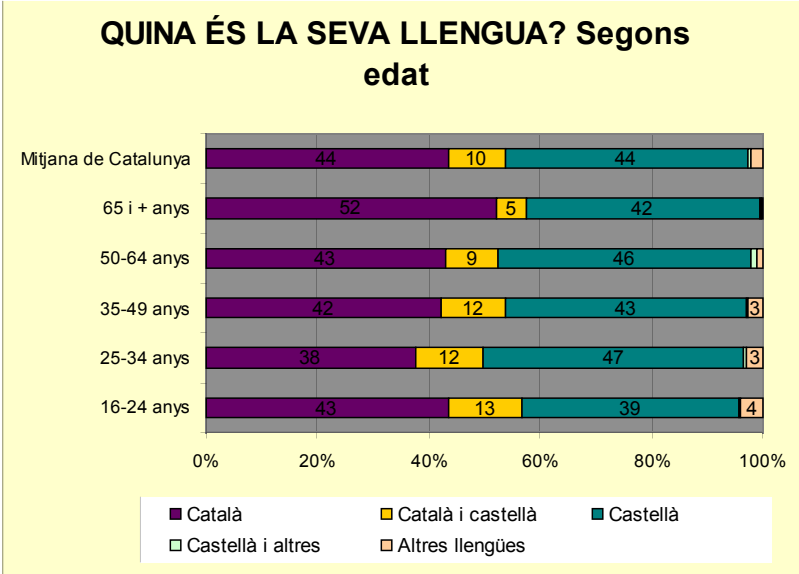
En el cas de la classe mitjana propietària, empresaris i autònoms, el petit percentatge que presenta no ens permet diferenciar-hi subgrups, però es fa evident que hi ha una segmentació interna important entre aquells que gaudeixen de millors condicions econòmiques davant dels que en tenen de pitjors.

Els grups socials i la llengua

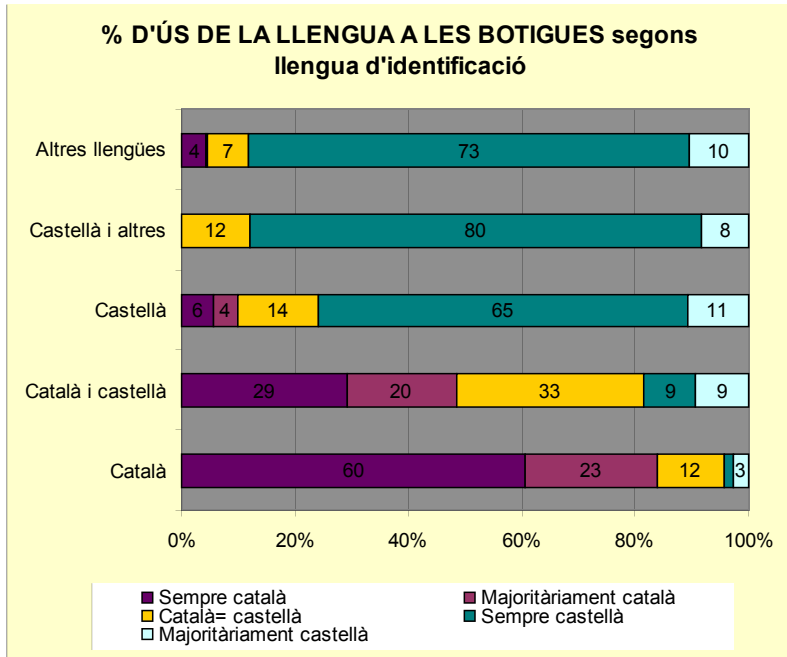
Una primera anàlisi de les variables lingüístiques del PaD ens mostren que a la nostra societat hi ha dos grans grups lingüístics: un que considera que la seva llengua és el català i un altre que considera que la seva llengua és el castellà; i un grup fronterer (10%) que considera que són ambdues llengües les seves. A més, en un 2% dels casos, i atesa l'evolució social en creixement, tenim un grup que no s'identifica ni amb la llengua catalana ni amb la castellana.



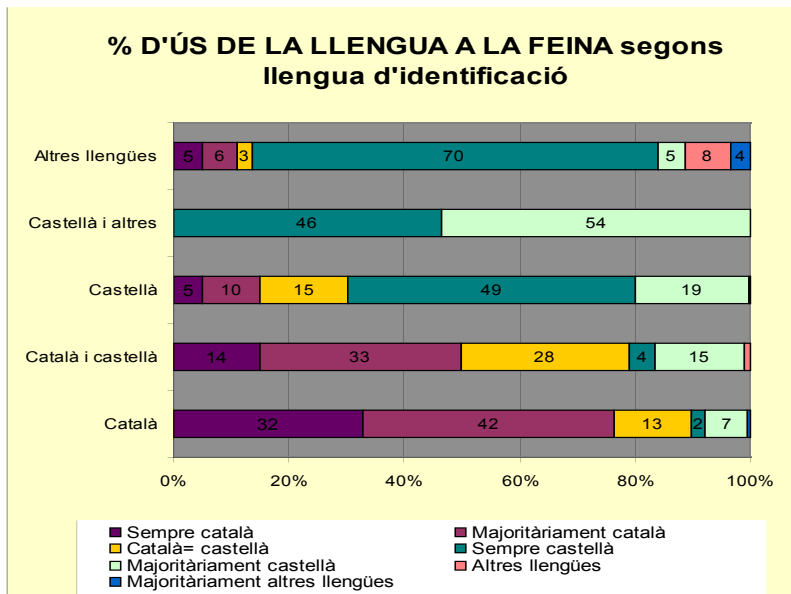
Aquests grups de llengua mostren diferències importants en relació amb l'edat dels individus i el lloc de residència. En els grups més joves augmenta la identificació bilingüe català-castellà i els casos d'altres llengües. Entre els més grans de 65 anys és majoritari el català, cosa que no succeeix en cap altre grup d'edat tret del dels més joves (16-34 anys).



En el territori, és la província de Barcelona on més es considera que la llengua és el castellà, seguida a distància per la província de Tarragona.



Respecte a la llengua d'ús trobem diferències importants, segons la llengua d'identificació de cada grup, en el comportament dels ciutadans segons l'àmbit social en la qual es realitza la interacció.



L'anàlisi de correspondències múltiples en l'àmbit de la llengua

A partir d'aquesta primera exploració de les dades sobre llengua del PaD es planteja l'anàlisi de correspondències múltiples amb les següents variables:

- Llengua en la qual es va realitzar l'entrevista
- Llengua de l'entrevistat, en resposta a la pregunta quina és la seva llengua
- Inici conversa amb un desconegut
- Ús de la llengua a l'escola, freqüència de percentatge de cada llengua
- Ús de la llengua a la feina, freqüència de percentatge de cada llengua
- Ús de la llengua a la llar, freqüència de percentatge de cada llengua
- Ús de la llengua a les botigues, freqüència de percentatge de cada llengua
- Ús de la llengua amb els veïns, freqüència de percentatge de cada llengua
- Ús de la llengua amb els amics, freqüència de percentatge de cada llengua

Els resultats de l'anàlisi de correspondències més la partició jeràrquica permeten identificar sis categories diferents de població caracteritzades tant per la llengua d'identificació com per l'ús de les diferents llengües en la interacció social, marcant amb especial rellevància les diferents situacions d'alternació de llengües que es produeixen contínuament a la nostra societat. Vegem en detall quins han estat els elements significatius en la definició de cada un dels grups.

Grup 1 (2,51%)

Bilingües de castellà i català més d'ús que d'identificació i especialment d'ús social. Només el 26% del grup considera que la seva llengua és tant el català com el castellà però utilitzen ambdues normalment amb amics, veïns i botigues (95%) i en canvi molt menys a la feina llar o en inici de conversa amb un desconegut (del 23% al 38%).

Grup 2 (22,95%)

Només el 26% del grup s'identifica amb català i castellà, però usen habitualment totes dues llengües. Utilitzen tant el català com el castellà en la interacció social, encara que hi ha més tendència a utilitzar el primer i a més a més també fan servir altres llengües (90%). El 44% usa català i castellà a les botigues, 36% amb els amics, 27% amb els veïns, 23% a la feina, 22% en inici de conversa amb un desconegut i 17% a la llar. Un 10% contesten que la llengua de l'entrevista els és indiferent i el 66% la contesten en català.

Grup 3 (12,52%)

Aquest és un grup que identifica el català com la seva llengua i es caracteritzen per usar-lo sempre a la llar i majoritàriament fora d'aquesta. El 92% del grup usa sempre o majoritàriament el català. En el 77% dels casos l'usen sempre a la llar i el 70% ho fan majoritàriament amb els veïns, 79% amb els amics, 67% a les botigues, 87% inici conversa, 43% a la feina, 12% a l'escola i 9% a la llar.

Grup 4 (24,78%)

La seva llengua d'identitat és el català i es caracteritza pel fet d'usar-la sempre que poden. El 93% diuen que la seva llengua és el català, el 96% han fet l'entrevista en català, el 99% usa sempre o majoritàriament el català en tots els àmbits socials analitzats. El 90% usa sempre el català amb els veïns, el 86% amb els amics, el 92% a les botigues, el 95% a la llar, el 26% a la feina i majoritàriament el 18% i el 9% a l'escola.

Grup 5 (2,41%)

La seva llengua d'identificació no és ni el català ni el castellà (84%), però fan servir el castellà com a llengua d'ús. El 89% han fet l'entrevista en castellà, el 88% inicien la conversa amb un desconegut en castellà, el 73% parla castellà a les botigues, el 44% ho fa a la feina i el 79% ho fa amb els veïns.

Grup 6 (34,83%)

Tant la seva llengua d'identificació com la d'ús és el castellà. El 94% diu que la seva llengua és el castellà i el 99% usa sempre o majoritàriament el castellà en la interacció social. El 88% ha fet l'entrevista en castellà, el 95% inicia la conversa amb un desconegut en castellà, el 83% parla a les botigues en castellà i l'11% ho fa majoritàriament, el 32% parla a la feina en castellà i l'11% ho fa majoritàriament, el 96% parla amb els veïns en castellà, el 90% amb els amics i el 94% parla sempre en castellà a la llar.

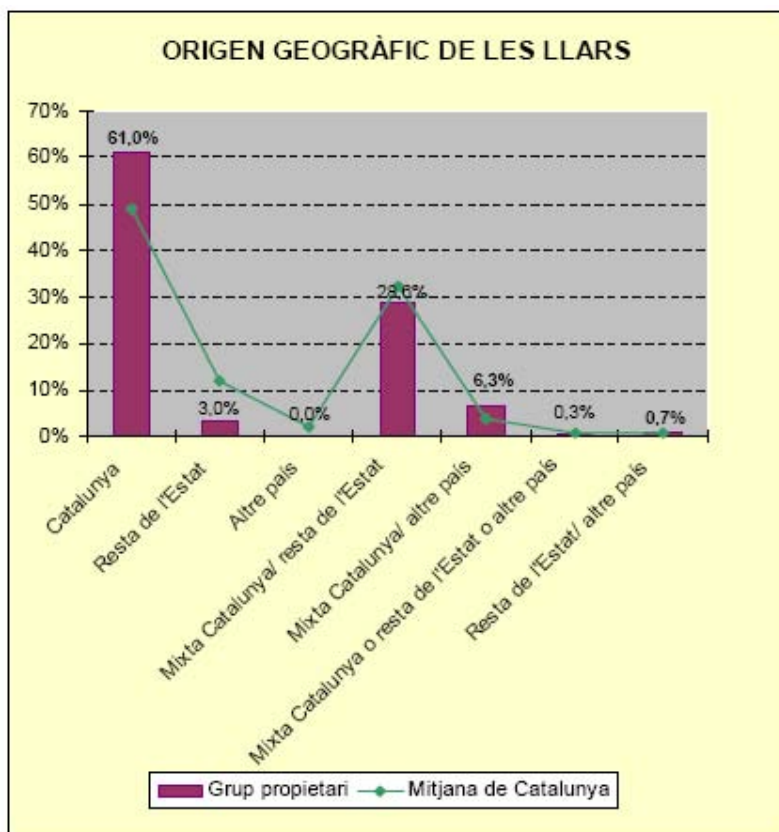
La relació que s'estableix entre aquesta tipologia de grups de llengua i la definició dels grups socials en la realització de l'anàlisi de correspondències final queda detallat en el quadre que apareix a continuació.

Grups de classe social	Tipologia de llengua El que els defineix	Tipologia de llengua El que no els defineix
Classe mitjana propietària		Grup 6 (23,6%)
Classe mitjana assalariada en posicionament	Grup 2 (44%)	Grup 1 (0,79%) Grup 6 (31,4%) Grup 4 (9,5%)
Classe mitjana assalariada posicionada	Grup 6 (55%)	Grup 4 (19,5%) Grup 2 (12,7%)
Classe mitjana assalariada benestant	Grup 3 (26%) Grup 2 (29%)	Grup 6 (5,7%)
Classe treballadora precària	Grup 5 (13%)	Grup 2 (16%)
Classe treballadora formalitzada	Grup 6 (50%) Grup 2 (34%)	Grup 3 (4%) Grup 4 (6,4%)
Classe treballadora vella i pobre	Grup 6 (72%)	Grup 4 (14,5%) Grup 2 (10%) Grup 3 (1,55%)
Classe treballadora vella amb patrimoni	Grup 4 (39%)	Grup 1 (4,4%) Grup 5 (0,67%) Grup 2 (10,8%)

La caracterització dels grups socials

Comencem pels grups de classe mitjana. El grup de classe mitjana propietària (9,18%) no es caracteritza significativament per cap grup de la tipologia de llengua, però sí que podem dir que no es caracteritza pel grup 6, és a dir per aquells que s'identifiquen amb la llengua castellana i l'usen sempre o majoritàriament en la interacció social.

Aquest és un grup amb un origen majoritàriament català: el 61% són nascuts a Catalunya i de les llars en el 59% dels casos ho són els dos membres de la parella. Però és un grup on també tenen cabuda altres orígens geogràfics tant de la resta de l'Estat com de fora: hi ha un 9% de llars amb parelles nascudes fora d'Espanya i un 6% de parelles mixtes Catalunya i un altre país. El 53% diu que la seva llengua és el català i el 12% són bilingües català/castellà. De fet, acostumen a usar més el català (43%), però l'ús del català perd terreny a la feina en favor del bilingüisme.



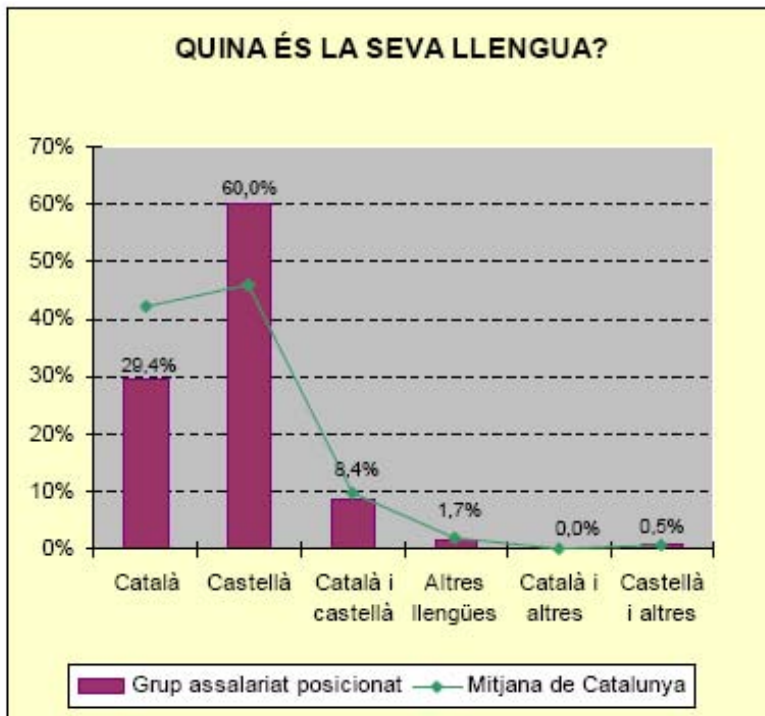
Dins la classe mitjana assalariada els tres grups identificats mostren trets característics que els diferencien. La llengua n'és un. El grup de classe mitjana assalariada en posicionament es caracteritza pel grup 2 de la tipologia de llengua, que representa el 44% del grup. Són aquells que no només usen el català i el castellà, sinó també altres llengües, i, encara que usen indistintament català i castellà, és lleugerament superior l'ús del català. La classe mitjana assalariada benestant també es caracteritza per aquest grup, però en un grau menor (29%) i alhora també es defineix pel grup

3 de la tipologia de llengua (26%). És a dir, aquells que identifiquen el català com la seva llengua i l'utilitzen sempre a la llar i majoritàriament a fora d'aquesta. En canvi, el grup de classe mitjana posicionada es caracteritza pel grup 6, que representa el 55% dels casos, és a dir, per identificar-se amb el castellà i usar-lo sempre o majoritàriament en la interacció social.

En el grup dels assalariats en posicionament (14,53%), és a dir els més joves, són majoritàriament nascuts a Catalunya (59%) i el 58% de les parelles nucli de la llar també compleixen aquesta condició. La resta són parelles mixtes (28%) i també hi ha un 10% de parelles nascudes fora de l'Estat. Encara que són majoritàriament nascuts a Catalunya, la seva llengua no és majoritàriament el català: és el castellà (50%) i el català i el castellà de manera igual en el 13% dels casos. De fet, conviuen en aquest grup tres realitats lingüístiques diferents: uns que parlen majoritàriament en totes les situacions (amics, feina, llar, botigues) en català, uns en castellà i uns que usen igual el castellà que el català.

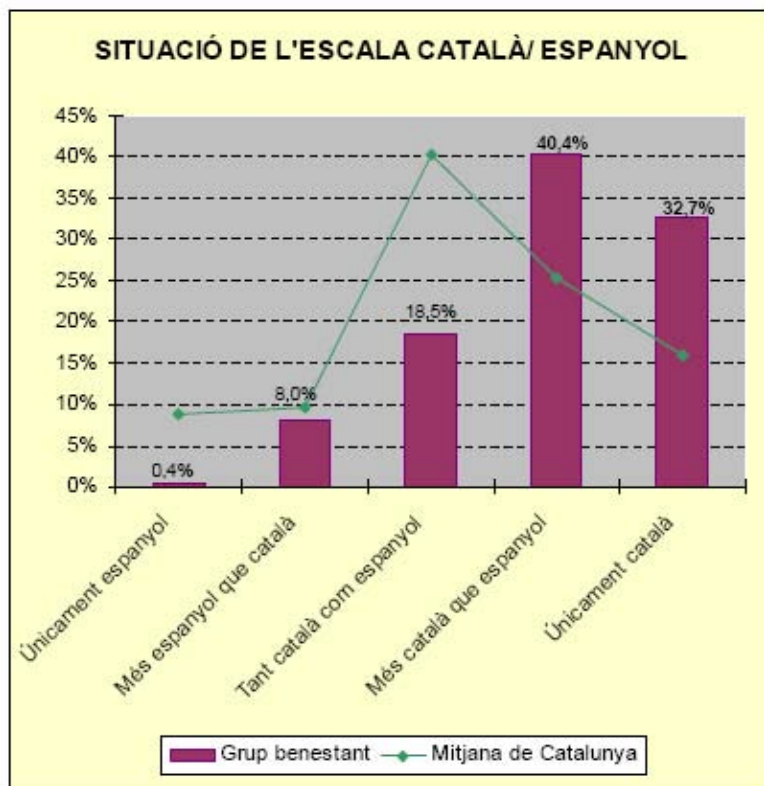
En el posicionament nacionalista del grup sembla que retrobem els tres grups de comportament lingüístic que hem descrit més amunt, però amb pesos diferents: el 13% se senten més espanyols que catalans, el 45% igual d'espanyols que de catalans i el 28% més catalans que espanyols.

En el segon grup de la classe mitjana assalariada, el grup assalariat posicionat (13,59%), hi predominen les llars amb parelles nascudes a la resta d'Espanya (28%) o parelles mixtes de nascuts a Catalunya i a la resta de l'Estat (25%). El 60% del grup diu que la seva llengua és el castellà, parlen habitualment per no dir sempre en castellà, especialment amb el grup d'amics, amb els veïns i a la llar. En el tema del nacionalisme, la meitat del grup es considera tant català com espanyol, i destaca un 13% que només es considera espanyol.



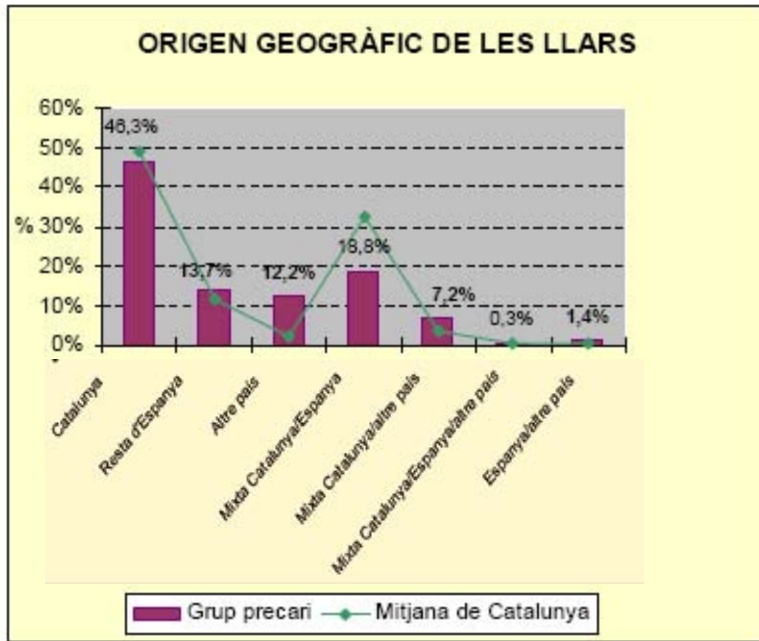
Per acabar la classe mitjana assalariada queda el grup benestant (12,95%), que és el grup més català en origen i en llengua. El 68% són nascuts a Catalunya i en el 70% de les llars ho són els dos

membres de la parella. Han fet el nombre més petit (8%) d'entrevistes en castellà. Parlen sempre o majoritàriament en català a la feina, amb els veïns parlen una mica més en castellà i a les botigues parlen majoritàriament en català. Una part important del grup (28%) també usa habitualment altres llengües que no són ni el català ni el castellà. Són, a més, els que mostren una consciència nacional més clarament catalana: el 40% es consideren més catalans que espanyols i el 33% només catalans.



La classe treballadora tampoc es presenta com un bloc uniforme. Hi trobem quatre grups diferenciats pel cicle vital que marca les diferents etapes de relació amb el mercat laboral i amb la vida reproductiva de les persones. Dels quatre grups socials que la formen hi ha dos grups molt envellits, amb la majoria dels seus membres jubilats, però amb diferent situació econòmica i lloc de residència. Dels altres dos, hi ha un grup fortament marcat per l'etapa reproductiva dels seus integrants, i un altre grup, el primer que comentarem, en un clar procés de posicionament en el mercat laboral i en la societat. Pel que fa a la llengua, també mostren diferències importants. El grup precari es caracteritza pel grup 5 de la tipologia, que representa el 13% dels casos, que són aquells que fan servir el castellà però la seva llengua no és ni el castellà ni el català. El grup de classe treballadora formalitzat es caracteritza per dos grups de la tipologia de llengua; el 6,50% dels casos; i el 2,34% dels casos. El primer és el que es defineix per tenir el castellà com a llengua i usar-la sempre o majoritàriament. El segon, en canvi, són aquells que usen català i castellà i d'altres llengües. El grup vell i pobre de la classe treballadora també es caracteritza pel grup 6 de la tipologia de llengua en el 72% dels casos, i en canvi, el grup vell amb patrimoni ho fa pel grup 4 de llengua (39%). Aquest és l'únic grup de classe social que té com a definició el grup 4 de la tipologia, és a dir aquells que s'identifiquen amb el català i l'usen sempre que poden.

L'origen geogràfic és el que fa més singular el grup precari (9,71%) de la classe treballadora. Aquest no és pas un grup d'immigrants, ni tots els immigrants són en aquest grup (54,5%), però hi ha un 12% de les llars amb tots els membres nascuts fora de l'Estat espanyol. La xifra és més alta encara si observem la composició del nucli de la llar: de llars en les quals els dos membres de la parella són de fora de l'Estat n'hi ha un 21%. I l'última dada que situa la procedència de bona part del col·lectiu és que els pares del 30% dels entrevistats d'aquest grup viuen en un altre país.



S'ha d'entendre, doncs, que la situació i el comportament lingüístic siguin molt diversos: la llengua del 45% del grup és el català, el 38% el castellà i la de l'11% altres llengües que no són ni el català ni el castellà. Però la llengua vehicular és el castellà: el grup en conjunt usa majoritàriament el castellà a la feina, amb els amics tant el castellà com altres llengües i amb els veïns català i altres llengües. Només usa sempre o majoritàriament el català el 36% del grup.

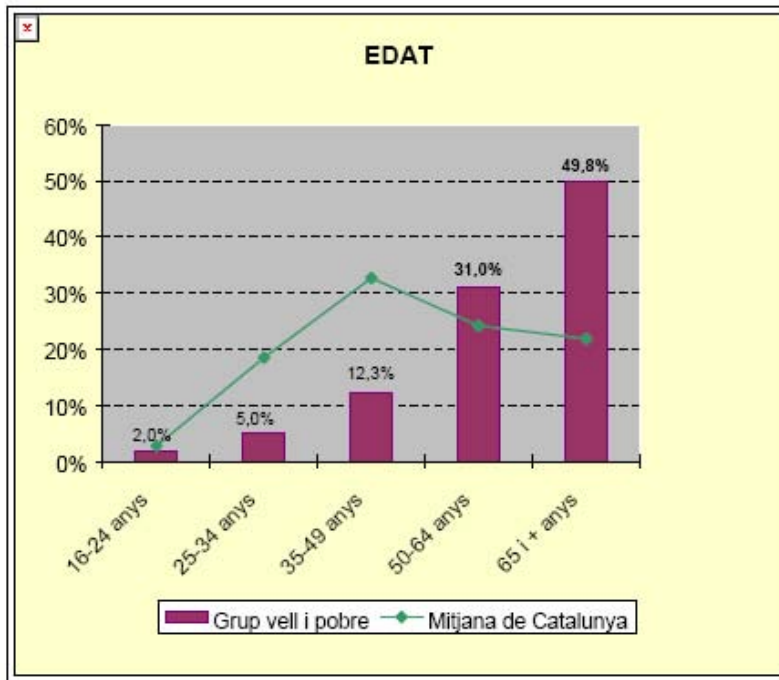
La posició respecte del nacionalisme repeteix aquesta dualitat: malgrat la resposta majoritària, és igual de català que d'espanyol, els percentatges més elevats en comparació als altres grups són més català que espanyol (26%), només català (18%) i només espanyol (10%).



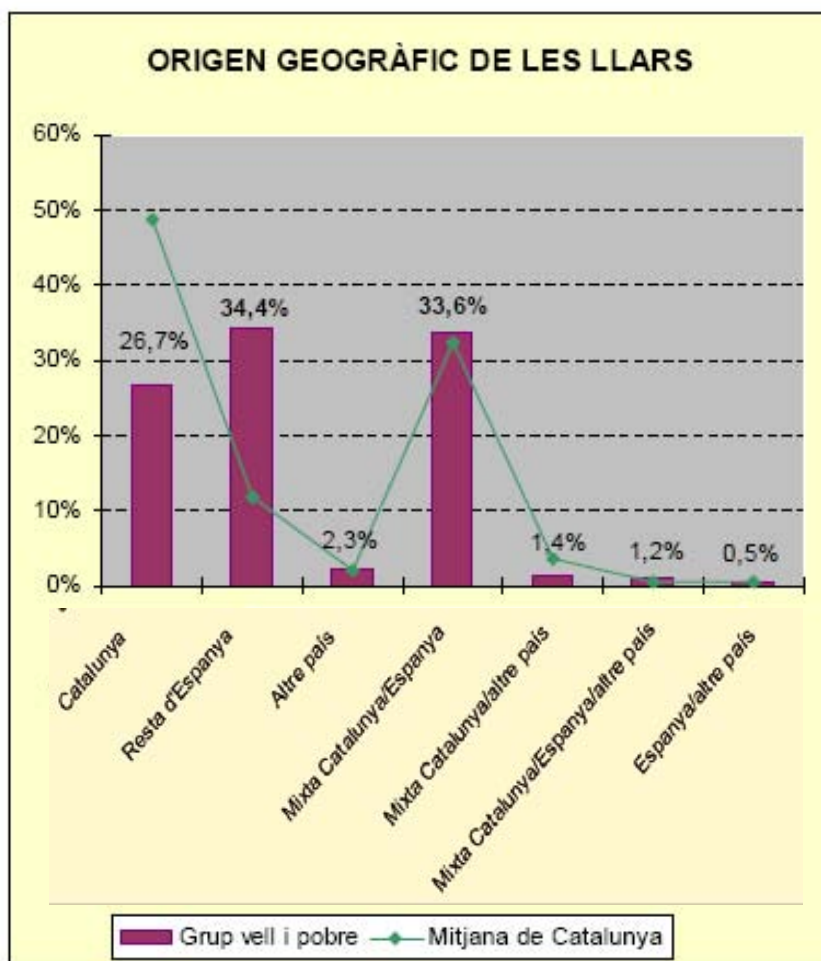
En el segon grup de classe treballadora, l'origen geogràfic del grup formalitzat (8,24%) no és especialment significatiu –a la majoria de llars tots els membres són nascuts a Catalunya i gairebé a la meitat ho són els dos membres de la parella (49%)–; tret que hi ha un 28% de parelles mixtes amb un membre nascut a Catalunya i un a la resta de l'Estat. També hi ha un 7% de parelles nascudes

fora d'Espanya. Malgrat que predominen els nascuts a Catalunya, la seva llengua és el castellà (62%) —el 61% ha fet l'entrevista en castellà— i només es consideren bilingües l'11%. Els percentatges de llar són paral·lels als individuals: al 63% de les llars sempre parlen en castellà i al 12% castellà i català. Alhora, quan els demanem quina és la llengua que fan servir a les botigues, amb els amics, amb els veïns i a la feina la resposta majoritària també és el castellà, tot i que sobretot a la feina i amb els amics augmenten els casos dels que parlen tant en castellà com en català (20%). Malgrat que gairebé la meitat del grup es consideren tant catalans com espanyols, l'11% es defineix només espanyol i el 15% més espanyol que català.

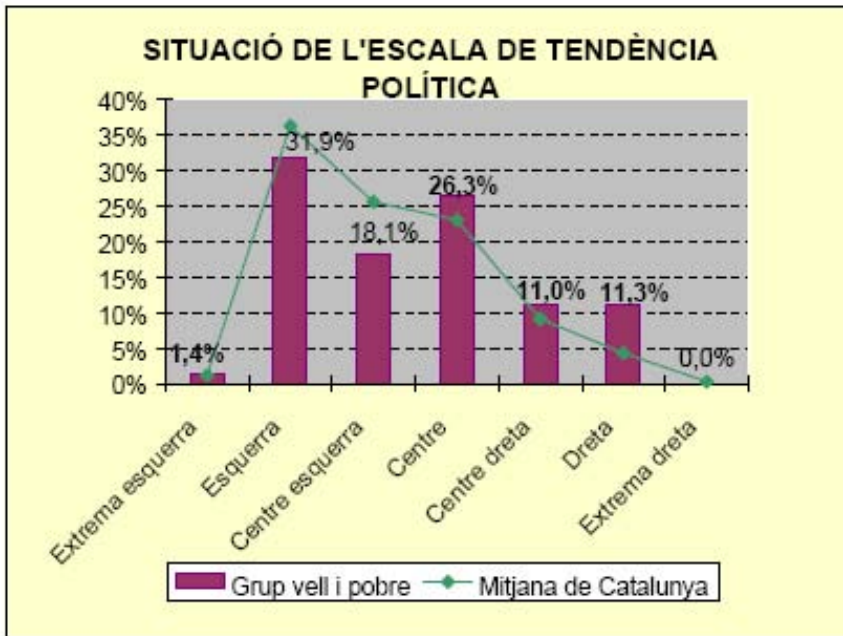
El grup vell i pobre (9,79%) està fortament marcat per les onades migratòries de mitjan segle passat: el 64% del grup ha nascut a Espanya, en el 56% de les llars els dos membres hi han nascut i a una de cada 3 llars ho han fet tots els seus membres.



El 75% dels grups parla en castellà tant a dins com a fora de casa (amics, veïns, botigues, desconeguts). Torna a ser un grup que resideix a la província de Barcelona (83%) majoritàriament a ciutats mitjanes i a Barcelona ciutat. I es consideren més espanyols que no pas catalans: el 20% només espanyols, el 15% més espanyol que catalans.



L'últim grup de classe treballadora, el vell amb patrimoni (22%) té un component menys castellà que el grup anterior, però la presència dels vinguts de la resta de l'Estat és prou important: el 47% dels entrevistats ho són, i al 23% de les llars, tots els membres ho són, en una tercera part (34%) hi han nascut els dos membres del nucli. Però en canvi, a diferència del grup anterior, la llengua que hi predomina (51%) és la catalana. El 49% del grup parla sempre o majoritàriament en català –quan més canvien de llengua és amb els veïns i a la feina– i el 5% usen igual el català i el castellà.



Respecte del nacionalisme, segueixen el perfil general, però estan una mica més polaritzats als extrems: el 9% del grup es defineixen només espanyols i el 18% només catalans.

Conclusions

Els resultats de l'anàlisi duta a terme mostren la relació que s'estableix actualment entre els grups de classe social i la llengua. No hi ha grups socials identificats en exclusivitat ni per una llengua ni per un model d'interrelació lingüística, però això no vol dir que no s'estableixi una relació entre la posició social i la llengua perquè sí que hi ha grups lingüístics i usos de la llengua diferenciats per grups socials.

La llengua castellana és molt més característica de la classe treballadora que no pas de la classe mitjana, i la llengua catalana a l'inrevés. De fet, l'àmbit de la llengua ha estat un element més que ha definit les diferències entre les classes socials, junt amb altres aspectes de les condicions de vida, usos i comportaments dels individus de la mostra. Però dit això, la fotografia és molt més complexa, perquè els diversos grups identificats dins de cada classe social tenen característiques diferents i alhora no tot el grup comparteix les mateixes.

Així doncs, la classe treballadora s'identifica majoritàriament amb el castellà, però el grup anomenat classe treballadora vella amb patrimoni és el grup "més" català de tots. És a dir, l'únic que té com a característic el grup 4 de llengua.

Ahora dins de la classe mitjana –que en conjunt té el català o el bilingüisme com a trets característics–, el grup que hem anomenat classe mitjana posicionada no compleix aquest comportament, ans al contrari, té com a element característic el grup 6 de llengua, és a dir tenir el castellà tant com a llengua d'identificació com llengua d'ús.

Bibliografia

- BATISTA, J. M.; SUREDA, J. (1987). "Análisis de correspondencias y técnicas de clasificación: su interés para la investigación en las ciencias sociales y del comportamiento". *Infancia y Aprendizaje*, núm. 39-40, pp. 171-186.
- BAXTER, J. et al. (1991). *Class analysis and Contemporary Australia*. Melbourne: The Macmillan Company of Australia Pty. Ltd.
- BECHHOFFER, F.; ELLIOTT, B. (1983). *The Petite Bourgeoisie. Comparative Studies of the Uneasy Stratum*. London: The MacMillan Press Ltd.
- BENZECRI, J. P. (1973). *L'analyse des données*. París: Dunot.
- BERGER, P. L. (1991). *La revolución capitalista. Cincuenta proposiciones sobre la prosperidad, la igualdad y la libertad*. Barcelona: Ediciones Península.
- BOURDIEU, P. (1988). *La distinción. Criterios y bases sociales del gusto*. Madrid: Taurus.
- BOTTOMORE, T. (1991). *Classes in Modern Society*. London: Harper Collins Academic.
- BURROWS, R. i MARSH, C. (1992). *Consumption and Class. Divisions and Change*. New York: St. Martin's Press.
- CACHÓN, L. (1989). *¿Movilidad social o trayectorias de clase?. Elementos para una crítica de la sociología de la movilidad social*. Madrid: Centro de Investigaciones Sociológicas i Siglo Veintiuno de España. Editores s.a., Colección Monografías, núm. 100.
- CAÍNZOS, M. A. (1989). "Clases, intereses y actores sociales: un debate postmarxista". *Revista Española de Investigaciones Sociológicas*, núm. 46, pp. 81-100.
- CALVERT, P. (1982). *The Concept of Class*. Nova York: St. Martin's Press.
- CARABAÑA, J. i de FRANCISCO, A. (comps) (1992). *Teorías contemporáneas de las clases sociales*. Madrid: Zona Abierta, núm. 59-60.
- CARCHEDI, G. (1975). "Reproduction of social classes at the level of production relations". *Economy and Society*, vol. 4, núm. 4, pp. 361-417.
- CARCHEDI, G. (1987). *Class Analysis and Social Research*. Oxford: Basil Blackwell Ltd.
- COMUNIDAD DE MADRID (1992). *Clases sociales: estudio comparativo de España y la Comunidad de Madrid 1991*. Carabaña, J. (dir.) i González, J. J. (coord.), Madrid, Consejería de Economía de la Comunidad de Madrid.
- CORNEJO, J.M. (1988). *Técnicas de investigación social: el análisis de correspondencias*. Barcelona: Promociones y Publicaciones Universitarias.
- CREDOC (1984). *Le système d'enquetes sur les conditions de vie et aspirations des français 1978-1982*. Rapport preliminaire Phase V, i Rapport final phase V. Paris, Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie.
- CROMPTON, R. (1993). *Class and Stratification. An Introduction to Current Debates*. Great Britain: Polity Press.
- MARSHALL, G. et al. (1989). *Social Class in Modern Britain*. London: Unwin Hyman Ltd.
- MIGUÉLEZ et al. (1997). *Desigualtat i canvi: l'estructura social contemporània*. Barcelona. Proa.

- RUNCIMAN, W. G. (1990). "How many Classes are there in Contemporary British Society?". *Sociology*, vol. 24, núm. 3, pp. 377-396.
- SCASE, R. (1992). *CLASS*. Buckingham. Open University Press
- SÁNCHEZ CARRIÓN, (ed.) (1984). *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales*. Madrid: Centro de Investigaciones Sociológicas.
- SÁNCHEZ, C. (1994). *La definició dels grups socials a la regió metropolitana de Barcelona. Un problema teòric i metodològic*. Tesis doctoral, Universitat Autònoma de Barcelona.
- SÁNCHEZ MIRET, C.; QUINTANA GARRIDO, N. (2005). "Les classes socials a Catalunya: desigualtats en les condicions de vida dels grups socials". A: D.A. *Estructura social i desigualtats a Catalunya*, Volum 1 (2005) Fundació Jaume Bofill.
- SUBIRATS, M., SÁNCHEZ, C. i DOMÍNGUEZ, M. (1992). *Grups i classes socials a la Regió Metropolitana de Barcelona. Enquesta de la Regió Metropolitana de Barcelona 1990. Condicions de vida i hàbits de la població*. Vol. 4, Informe general. Direcció general: Marina Subirats, direcció metodològica: Carlos Lozares. Barcelona, Institut d'Estudis Metropo
- VAN PARIJS, P. (1992). "Una revolución en la teoría de las clases". *Zona Abierta*, núm. 59-60, pp. 187-228.
- WRIGHT, E. O. (1989). *Classes*. Londres-Nova York: Verso.
- WRIGHT, E. O. (ed.) (1989). *The Debate on Classes*. Londres-Nova York: Verso.
- WRIGHT, E. O. (1992). "Reflexionando, una vez más, sobre el concepto de estructura de clases". *Zona Abierta*, núm. 59-60, pp. 17-126.
- WRIGHT, E. O. et al. (1992). "The Non-effects of Class on the Gender Division of Labor in the Home: A Comparative Study of Sweden and the united States". *Gender and Society*, vol. 6, núm. 2, pp. 252-282.

Mesurem els mots? L'anàlisi estadística de textos

Mónica Bécue-Bertaut

Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

Resumen

En las encuestas por cuestionario, es usual introducir preguntas abiertas. En los estudios de consumo, los expertos y/o consumidores evalúan los productos mediante una puntuación y, conjuntamente, un comentario libre. El análisis estadístico de este tipo de datos, llamados datos mixtos, requiere de métodos específicos. Proponemos aquí emplear una metodología que combina el análisis factorial múltiple clásico y el análisis factorial múltiple para tablas de contingencia. Esta extensión del AFM permite tratar globalmente tablas léxicas, creadas a partir de las respuestas textuales, y tablas cuantitativas y categóricas.

Presentamos la metodología apoyándonos en datos recogidos en la evaluación de dos conjuntos de vinos. El análisis sensorial constituye en efecto un área de aplicación privilegiada, dado que los expertos y consumidores desean frecuentemente complementar los clásicos perfiles sensoriales con descripciones libres que traducen más fielmente sus percepciones.

La metodología se puede emplear también en el tratamiento de datos de encuesta, cuando se aborda una problemática con una batería de preguntas cerradas y de preguntas abiertas.

1. Introducció

Conocer las opiniones de los usuarios y/o consumidores es un objetivo crucial en los estudios de mercados. En ciertos casos, dichas opiniones no se pueden transmitir mediante una puntuación o la selección de un ítem entre los propuestos y es necesario introducir preguntas abiertas. Los entrevistados expresan así su opinión simple o compleja, lacónica o rica en matices, a veces contradictoria, de manera espontánea.

En la industria agroalimentaria, está establecido que la calidad de los alimentos no se puede medir únicamente a partir de análisis químicos o físicos. Es importante coleccionar las percepciones de los consumidores y/o expertos mediante puntuaciones, apreciaciones sensoriales cualitativas y, más recientemente, comentarios libres. Los datos recogidos suelen ser voluminosos y requieren de la aplicación de métodos estadísticos, algunos específicos de este campo.

Presentamos aquí una metodología para el tratamiento de comentarios libres en estudios sensoriométricos y/o respuestas abiertas en encuestas.

La estructura de la exposición es la siguiente. En la sección 2, se recuerdan los principios básicos de codificación de los textos. En la sección 3, se presenta una metodología para el análisis simultáneo de varias tablas de frecuencia (Bécue & Pagès, 2004) y su aplicación a comentarios de cata en un estudio efectuado sobre vinos del Priorat. La sección 4 expone una metodología para

el análisis de tablas mixtas, con columnas-variables cuantitativas, categóricas y de tipo frecuencia y su aplicación al estudio de un conjunto de vinos, descritos mediante un comentario libre y una puntuación. La sección 5 describe brevemente una investigación en curso, destinada a comparar la apreciación de unos mismos vinos por dos paneles de catadores expertos, uno catalán y uno francés. La sección 6 concluye la exposición.

2. Análisis estadístico de respuestas abiertas

2.1. Respuestas abiertas y transmisión de opiniones complejas

Las respuestas abiertas aportan una información específica (Lebart et al., 2000) y permiten abordar temas difíciles y complejos que requieren respuestas espontáneas. Dichas respuestas no se pueden enmarcar en los estrechos límites de una corta serie de ítems, frecuentemente propuestos en las encuestas de opinión.

La selección del vocabulario expresa una realidad subyacente compleja y difícil de aprehender por otro tipo de cuestionamiento. El análisis estadístico parte de las ocurrencias y co-ocurrencias de las distintas palabras y constituye una herramienta poderosa para acceder al contenido de los textos (Benzécri, 1981, Lebart, 2003; Murtagh, 2008).

2.2. Codificación de la información textual

Se considera un nuevo tipo de variable –llamada variable textual– codificada en una tabla individuos×palabras (o tabla léxica), construida mediante el recuento de las diferentes palabras en las respuestas individuales. Esta estructura más compleja que la estructura inducida por una variable cuantitativa (una columna) o una variable categórica (tantas columnas como categorías) resulta de la variabilidad del lenguaje que es también su riqueza.

Identificar las unidades léxicas requiere un cuidadoso pretratamiento (Labbé, 1990; Muller, 1977-1992; Salton & MacGill, 1983) que:

- Corrige las faltas de ortografía.
- Opera una lematización que convierte la forma gráfica de cada ocurrencia en su forma estandarizada llamada lema (o voz del diccionario: infinitivo para los verbos, singular para los sustantivos, masculino singular para los adjetivos, etc.) e identifica la categoría gramatical de cada ocurrencia. Seguimos aquí la norma lexicométrica introducido por Muller (1977-1992) y completada por Labbé (1990) que limita las unidades léxicas compuestas por varias ocurrencias a las locuciones fuertemente establecidas en la lengua. A pesar de que ciertos lemas puedan modificar su significado en función del contexto –en particular, los adjetivos pueden estar fuertemente influenciados por un adverbio de cantidad–, se conserva los distintos lemas como unidades separadas. En lo que sigue, empleamos el término genérico *palabra* como sinónimo de *lema*.
- Define la stoplist, o lista de palabras consideradas no útiles en el estudio en curso. En general, dicha lista contiene las preposiciones, artículos, pronombres, conjunciones y adjetivos posesivos y demostrativos.
- Selecciona un umbral de frecuencia, dado que la comparación entre respuestas sólo tiene

significado si las palabras tienen una frecuencia mínima (Lebart et. al., 2000). Una regla pragmática consiste en eliminar las palabras escogidas por menos de 2% de los individuos.

Así, se adopta una norma para el recuento lexicométrico que es estable, comprensible y reproducible: no varía durante el pretratamiento de un corpus de textos, ni de un corpus a otro, es fácil de entender por los usuarios y de aplicar en cada estudio y/o por cada usuario.

El pretratamiento ofrece una codificación transparente de las respuestas, operado por el analista mediante un proceso explícito e igualmente aplicado a todas las respuestas.

Obviamente, la significación de algunas palabras puede variar según el contexto. Como lo ha estudiado Lehrer (1975), la potencia del lenguaje proviene de su labilidad y el significado impreciso de las palabras no es sino una virtud. El *perfil léxico* se debe aprehender globalmente para tomar en cuenta el contexto de las palabras.

Las palabras conservadas constituyen las columnas de la tabla léxica, que es una tabla de frecuencias particular. La suma de una fila corresponde a la longitud conservada de la correspondiente respuesta mientras que la suma de una columna es la frecuencia total de la palabra que la encabeza.

3. Hablemos de vinos

3.1. Datos y problemática

En 2003, la agrupación de viticultores del Priorat, en colaboración con INCAVI, quiso experimentar la contribución de los comentarios libres como herramienta de evaluación de los vinos. Así, 31 catadores —ellos mismos viticultores, productores de vinos o enólogos de las bodegas— degustaron 43 vinos provenientes del Priorat. El protocolo de la sesión contemplaba sólo una descripción libre de los vinos, aunque en el marco de una ficha muy precisa (Figura 1). Así, para cada vino, los cataadores describían cada uno de los aspectos clásicos del vino (visual, olfativo y gustativo) mediante las palabras asociadas a cada uno de los atributos mencionados en la ficha.

FITXA	ASPECTE VISUAL		EXAMEN OLFACTIU						EXAMEN GUSTATIU						A QUIN POBLE PERTANY
NÚM CATADOR	COLOR	INTENSITAT	FRANQUESA	INTENSITAT	PERFIL AROMÀTIC	TIPUS D'AROMA	CARÀCTER	QUALITAT	ESTRUCTURA TEXTURA	TANINS	VOLUM	RETRONASAL	POSTGUST	ARMONIA EQUILIBRI	
1	Violeta	media	si	media		?		bona	estruc Meda-baja	tanino dulce	medio	vainilla	làctico	mal equilibri	
												làctico		poco aroma	
2	vermellos	intens	si	intens		fruits madurs negres	ben integrat	correcte	complex	elegants	ample	llarg	agradable	correcte	
						espècies			madurs						
						tancat	complex moderat		potent	secants		fruits	mij secant	no	Gratallops
3	morat	molt intens	si	poc intens		fruits negres						fruits		massa secant	
				2/5		anis, fruits									

Figura 1. Ficha de comentarios asociados a un vino.

En su momento, no se trató dicha información sino mediante una lectura clásica. Jordi Escayola Mansilla emprendió la tarea de analizar estos datos para su proyecto de fin de carrera de la Diplomatura de Estadística (UPC) (Escayola, 2008).

3.2. Estructura de la tabla múltiple construida

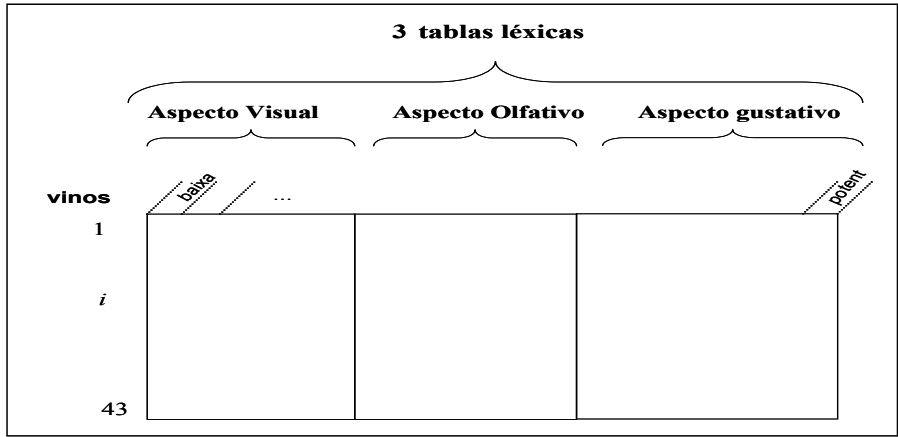


Figura 2. Tablas léxicas resultando de la codificación de las fichas

Se ha conservado sólo la división de las descripciones libres en los 3 aspectos visual, olfativo y gustativo, perdiendo la división más fina en atributos. Así, se consideran 3 variables textuales. A cada variable textual, corresponde un grupo de columnas-palabra.

Para cada uno de los 3 aspectos, se identifican las distintas palabras empleadas para describirlo y se recuenta su frecuencia. Además de lematizar, se reagrupan los sinónimos (por ejemplo, *vermell* y *vermellós* se reagrupan bajo el nombre genérico *vermell*). Después, se conservan las palabras empleadas al menos 4 veces para caracterizar un aspecto.

Así, se obtiene una tabla de frecuencias múltiple con, en fila, los 43 vinos y, en columna, 322 columnas-palabras divididas en 3 grupos (de, respectivamente, 39, 125 y 158 palabras) correspondientes a los 3 aspectos. Una misma palabra puede encabezar varias columnas si se ha empleado para describir diferentes aspectos. Para esto, la etiqueta de una palabra viene precedido de V, O o G según el aspecto que caracteriza.

3.3 Análisis de la tabla léxica múltiple

El análisis de correspondencias (AC) es el método en ejes principales de referencia para el análisis de una tabla de contingencia o de frecuencia (Benzécri 1973, Escofier, 2003; Lebart et al. 2006; Escofier and Pagès, 1988-1998). Su aplicación al análisis textual es frecuente (Benzécri, 1981; Lebart et al. 2000; Murtagh, 2005, 2008).

Se pueden aplicar el AC, por separado, a las 3 tablas (Visión, Olfato, Gusto) y comparar las estructuras inducidas sobre los vinos por cada grupo de columnas, pero la comparación es un trabajo arduo y la síntesis de los resultados es compleja. Además, no se dispondría de la estructura "global", es decir inducida globalmente por los 3 aspectos.

El análisis factorial múltiple para tablas de contingencia (AFMTC) (Bécue & Pagès, 1999, 2004) es una herramienta idónea para analizar la tabla léxica múltiple. Dicho método es una extensión del análisis factorial múltiple (Escofier & Pagès, 1998) cuyos principios recordamos primero.

3.3.1. Breve presentación del AFM

El análisis factorial múltiple analiza una tabla múltiple en la cual un conjunto de individuos está descrito por J grupos de variables, cuantitativas o cualitativas. El AFM realiza un análisis en componentes principales (ACP) no estandarizado de la tabla yuxtapuesta, pero ponderando las variables de manera a equilibrar los distintos grupos. Por esto, el peso de las columnas-variables del

grupo j se divide por λ_j^j , primer valor propio obtenido en los análisis separados –ACP o análisis de correspondencias múltiples (ACM) según la naturaleza de las variables– de la subtabla j . Dicho método ofrece resultados:

- Análogos a los del ACP o del ACM; principalmente una representación global de las filas (individuos) y de las columnas (variables o categorías);
- Específicos de las tablas múltiples : principalmente la representación superpuesta de las estructuras inducidas por cada una de las subtablas sobre el conjunto de los individuos –llamadas estructuras parciales– y la representación de los factores obtenidos en los análisis separados.

3.3.2. AFM para tablas de contingencia múltiples (AFMTC)

Bécue & Pagès (1999, 2004) han propuesto una metodología para el análisis simultáneo de un conjunto de tablas de contingencia. Dicho método parte de los principios del análisis de correspondencias binarias intra-tablas (Benzécri, 1983 ; Escofier & Drouet 1983), generalizada por Cazes & Moreau (1991, 2000) con el nombre de análisis de correspondencias interno (ACI), y , así, toma en cuenta las diferencias entre los márgenes de las filas. Además, se adopta el enfoque del AFM para equilibrar la influencia de las diferentes tablas y para proporcionar gráficos específicos de la estructura en grupos de las columnas. La figura 3 muestra la tabla a analizar y precisa la notación empleada.

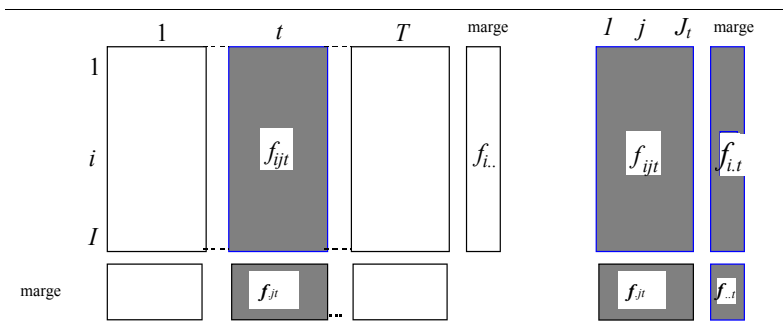


Figura 3. Tabla de contingencia múltiple y márgenes: notación.

En esta tabla, f_{ijt} : frecuencia relativa asociada a la fila i ($i=1, \dots, I$) y columna j ($j=1, \dots, J$) de la tabla t ($t=1, \dots, T$); un índice sustituido por un punto indica la suma sobre este índice.

Cada tabla léxica constituye una tabla de frecuencia, que se puede ver como una tabla de contingencia particular. Para introducirlos en un AFM, se sustituye cada uno por una subtabla de término general y_{ijt} (fila i , columna j , subtabla t):

$$y_{ijt} = \frac{f_{ijt} - \left(\frac{f_{i.t}}{f_{.t}} \right) \cdot f_{.j}}{f_{i.} \cdot f_{.j}} = \frac{1}{f_{i.}} \left[\frac{f_{ijt}}{f_{.j}} - \frac{f_{i.t}}{f_{.t}} \right]$$

El ACP de dicha tabla conduce a los resultados del ACI. Entre los corchetes, se tiene la desviación entre un perfil-columna y el perfil marginal de la tabla a la cual pertenece la columna.

En AFM, el equilibrio entre subtablas es obtenido por la surponderación de cada columna j por $1/\lambda_j^t$, inversa del primer valor propio del análisis separado de la subtabla j .

En el AFMTC, los análisis separados no son los AC usuales pero análisis de correspondencias con un peso impuesto, igual al margen en fila de la tabla yuxtapuesta (análisis llamados pseudo-separados). Así, las filas conservan el mismo peso en los distintos análisis. La deformación inherente a las sustitución de los AC separados por los AC pseudo-separados, mínima si los márgenes en fila difieren poco de una subtabla a otra, resulta del compromiso necesario a la comparación de tablas de contingencia con diferentes márgenes.

Después, se efectúa el análisis global que consiste en un ACP no estandarizado de la tabla yuxtaponiendo las subtablas de término general y_{ijt} definido antes con:

Pesos de las filas (y métrica en el espacio de las columnas): $\{f_{i.}; i = 1, \dots, I\}$, $f_{i.}$ es el peso relativo medio de los individuos calculado sobre la tabla yuxtaponiendo las tablas de contingencia;

Pesos de las columnas del grupo t (y métrica en el espacio de las filas): $\{f_{.jt}/\lambda_1^t; j = 1, \dots, J; t = 1, \dots, T\}$.

Este análisis ofrece resultados:

- Similares a los del AC aplicado a las tablas yuxtapuestas (principalmente, una representación global de los individuos-fila y de las palabras-columna).
- Específicos de las tablas múltiples, principalmente, la representación superpuesta de las estructuras inducidas sobre los individuos por cada grupo de columnas -estructuras parciales- y la representación de los factores derivados de los análisis separados.

La lectura de los resultados viene facilitada por numerosas ayudas a la interpretación del AFM.

3.4. Resultados

La tabla yuxtaponiendo las 3 tablas léxicas, correspondiente a los 3 aspectos descritos en la degustación, se ha analizado mediante el AFMTC descrito antes.

3.4.1. Visualización de los vinos-fila y de las palabras-columna sobre el primer plano factorial

Las figuras 4 y 5 muestran la representación de los vinos y de las palabras sobre el primer plano factorial. Debido al enfoque empleado, las palabras son representadas como lo son usualmente las variables en ACP. Otra opción consistiría en representarlas como en AC, en el centro de gravedad de los vinos en la descripción de los cuales se han empleado.

Los 3 grupos de palabras contribuyen de manera equilibrada a la inercia del primer eje. Dicha dirección de dispersión es común a los 3 grupos y no difiere mucho de los primeros ejes factoriales obtenidos en los análisis separados: el primer valor propio del análisis global (2.7) es próximo a su máximo (igual a 3 en este caso), que se obtiene cuando los ejes de las análisis separados son idénticos (Escofier & Pagès, 1998, p. 161).

Dicho eje contrasta las palabras indicando defectos de los vinos (*V-baixa*, *V-ataronjat*, *O-dolenta*, *G-animal*, etc.), a la izquierda del eje, con palabras asociadas con los mejores vinos, sobre todo relativas a las características olfativas y gustativas (*O-excel·lent*, *O-complexa*, *O-elegant*, *G-bona*, etc.), a la derecha del eje. Se puede interpretar como un eje de “calidad”.

El segundo eje no tiene una interpretación tan clara, aunque se puede ver que, también, refleja la calidad de los vinos.

Nos enfrentamos aquí a una dificultad: en el análisis textual la interpretación se apoya sobre la información suplementaria conocida, aquí ausente.

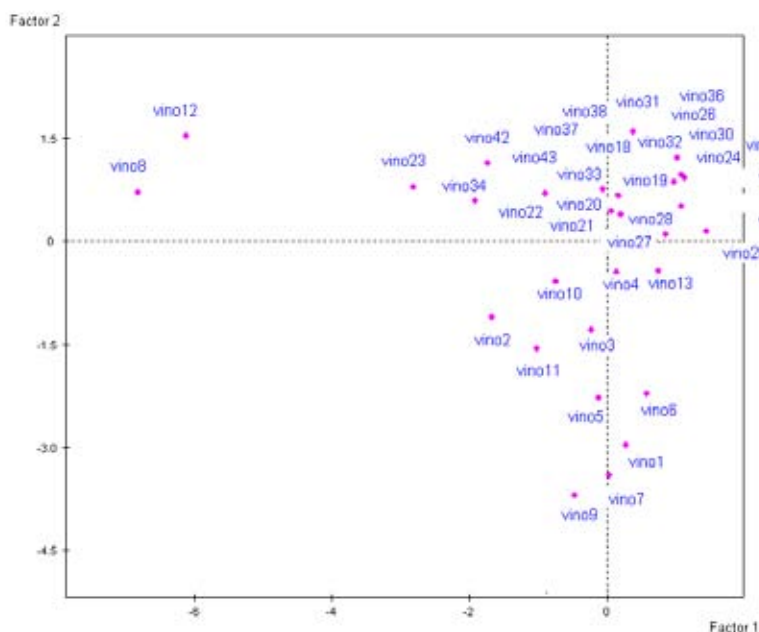


Figura 4. Los vinos sobre el primer plano factorial

Se han eliminado de esta representación algunos de los vinos en la zona central

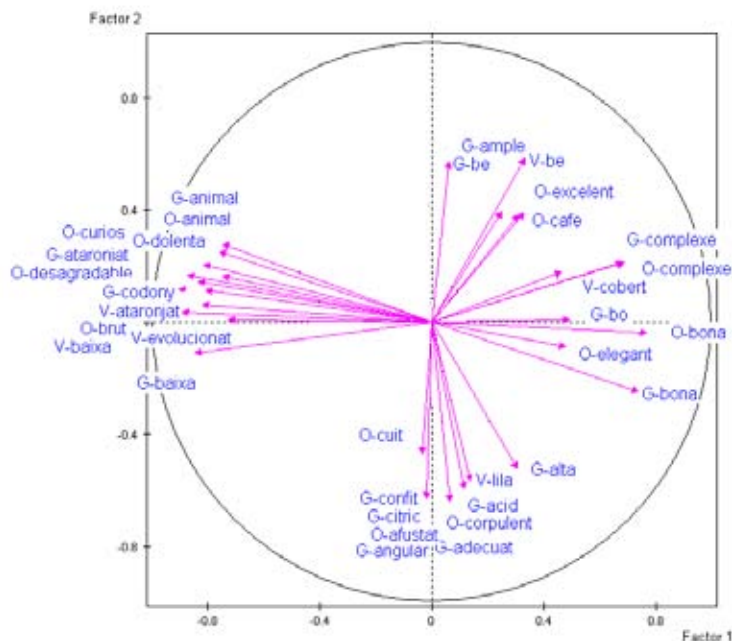


Figura 5. Representación de las palabras más contributivas sobre el primer plano factorial

3.4.2. Superposición de las representaciones parciales y globales

Es posible que un vino no reciba la misma apreciación según el aspecto considerado. Un vino juzgado excelente por sus características olfativas puede, por ejemplo, revelarse decepcionante en boca.

El AFMTC permite representar a cada individuo según el punto de vista global (punto global, como en la figura 4) pero también según los puntos de vista que corresponden a los 3 grupos de columna (puntos parciales) (Escofier & Pagès, 1998; Bécue & Pagès, 2008a).

La desviación entre los puntos parciales correspondientes a un mismo vino se interpretan según el significado dado a los ejes. Así, por ejemplo, el vino 8, de baja apreciación global, lo es sobre todo por sus características visuales mientras que el vino 12 tiene una apreciación más negativa desde el punto de vista olfativo y gustativo que desde sus características visuales. Los vinos destacados en la parte negativa del eje 2, ofrecen ciertas contradicciones entre los aspectos olfativo y gustativo.

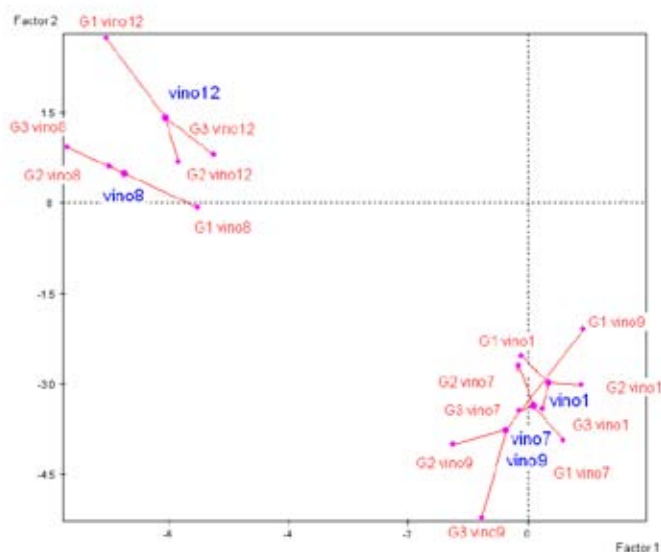


Figura 6. Representación global y parcial de los vinos 8, 12, 9, 7, 1 sobre el primer plano factorial

4. Utilización de información cerrada

En la sección anterior, hemos encontrado dificultades en interpretar los resultados debido a la ausencia de información suplementaria. Otro ejemplo, escogido también en el mundo del vino, permitirá mostrar la aportación de la dicha información.

4.1. Ejemplo: Una guía de vinos

El ejemplo trata de una guía dedicada a los vinos de Castilla y León (El Mundo, 2005) que presenta 522 vinos, entre los cuales 364 tintos. Cada vino está evaluado por comentarios de cata y una puntuación (sobre 100) que, para los vinos presentados en esta edición, varía entre 70 et 97. Además, está indicado el precio, la cepa y la añada. Después del pretratamiento de los comentarios, se conservan 222 palabras. La tabla léxica cruza los 364 tintos (filas) y las 222 palabras (columnas).

La tabla 1 muestra algunos comentarios. Se pone de manifiesto, y esto se verifica en el conjunto de la guía, que los vinos mejor puntuados reciben comentarios más extensos. Es un fenómeno clásico

en el análisis textual: en una encuesta por cuestionario, más interesado se muestra el entrevistado, más larga es su respuesta; si deseamos saber qué tema interesa más a un político, busquemos las frases más largas en sus intervenciones.

¿Cómo utiliza la información suplementaria, en particular la puntuación? Es el objeto de esta sección.

---- **Vino 53 (80) Mesoneros de Castilla-2003**

Tinto tempranillo joven limpio y típico, con fruta de hueso en la nariz. en boca los taninos resultan algo rasos.

---- **Vino 212 (85) Legaris-2001**

Tuestes, gominolas y buenos balsámicos marcan la intensidad media frutal de este crianza. En boca aparece muy lineal, con consistencia media; el retrogusto frutal todavía tapado por una madera algo rústica.

---- **Vino 30 (91) Tares P3-2001 premium**

Mucho terruño se detecta en el bouquet de este gran tinto; pólvora, sílex, pizarra, cascajo caliente con el contraste de tierra húmeda y mucha fruta madura de hueso. Concentrado, tacto graso sobre el paladar; impresionante viscosidad en la lengua, otra vez impresiones de tierra húmeda y pólvora en el largo final.

---- **Vino 314 (97) Vega Sicilia 'Único-1994**

Hay que realizar un ejercicio de disciplina gustativa de primer rango para describir este gran vino. El bouquet es fresco, bien armado de fruta roja que se ve potenciada por tintes de chocolates, tabacos, notas de sotobosque y una madera que se manifiesta pero que resulta difícil de localizar y menos de concretar. Tenemos el caso raro de un tinto que sale ileso del paso del tiempo sin lucir su armadura, que es la barrica. En boca joven, aunque ya tiene su cuerpo vigoroso y enérgico bastante ensamblado, con la excepción de algunos taninos saltamontes que quedan para domesticar. Largo y vibrante final que mezcla madurez con una notable finura fresca.

Tabla 1. Extracto de la guía de vinos

4.2. Introducció de les variables suplementaries en el anàlisi

4.2.1. Taula múltiple y notació

Columnas	Tabla léxica			Variable cuantitativa
Filas	1	j	J	Σ
1	Proporciones			
Vinos	i	$f_{ij} = \frac{n_j}{n}$		x_i
	I		$f_{i.}$	
Σ	$f_{.j}$		1	

Figure 7. Taula múltiple a analizar

La variable cuantitativa está estandarizada para los pesos inducidos

por el margen de la tabla léxica ($f_{i.}$)

- n_j : frecuencia de la palabra j en el comentario correspondiente al vino i ;
- n : longitud del corpus, es decir, número total de ocurrencias, sumado sobre el conjunto de los comentarios;

- $f_{ij} = \frac{n_j}{n}$: proporción de ocurrencias, sobre la longitud del corpus, de la palabra j en el comentario i sobre el vino i ;

- $f_{i.} = \sum_j f_{ij}$: proporción de ocurrencias correspondientes al vino i ;

- $f_{.j} = \sum_i f_{ij}$: proporción de ocurrencias de la palabra j ;

- x_i : puntuación dada al vino (*estandarizada*);

- γ_1^J : primer valor propio del AC aplicado a la tabla de frecuencias.

4.2.2. Posicionar las palabras sobre el eje de la nota

Una primera manera de estudiar la relación entre la puntuación (variable cuantitativa) y las palabras (columnas de la tabla léxica) consiste en emplear la “media ponderada” (weighted averaging, Orlaci, 1975). Cada palabra se posiciona mediante la media ponderada de las puntuaciones recibidas por los vinos cuyos comentarios contienen dicha palabra. La ponderación corresponde a la importancia relativa de la palabra en el comentario del correspondiente vino. Así, a la palabra j corresponde el valor $\bar{x}_j$, obtenido mediante el siguiente cálculo:

$$\bar{x}_j = \sum_{i=1}^I \frac{f_{ji}}{f_j} x_i$$

La figura 8 muestra la gráfica resultante. Además, se han posicionado los vinos con mejor puntuación. Esta figura ordena las palabras desde la más negativa, (en el contexto de una guía que sólo menciona vinos con una calidad aceptable) a la más positiva, desde *corto*, *correcto* y *herbáceo* a *enérgico*, *gran* y *denso*.

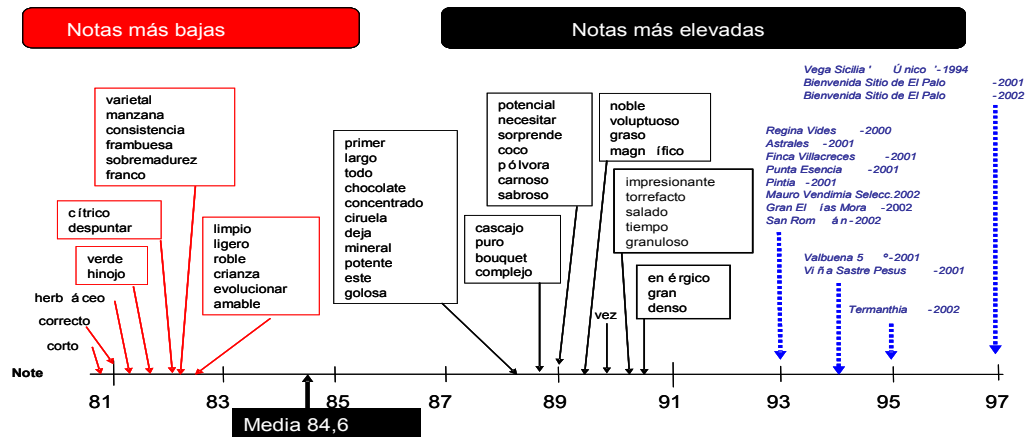


Figura 8. Posicionamiento de las palabras sobre el eje de las notas

4.2.3. Posicionar las palabras sobre el primer plano factorial proporcionado por el AC de la tabla léxica

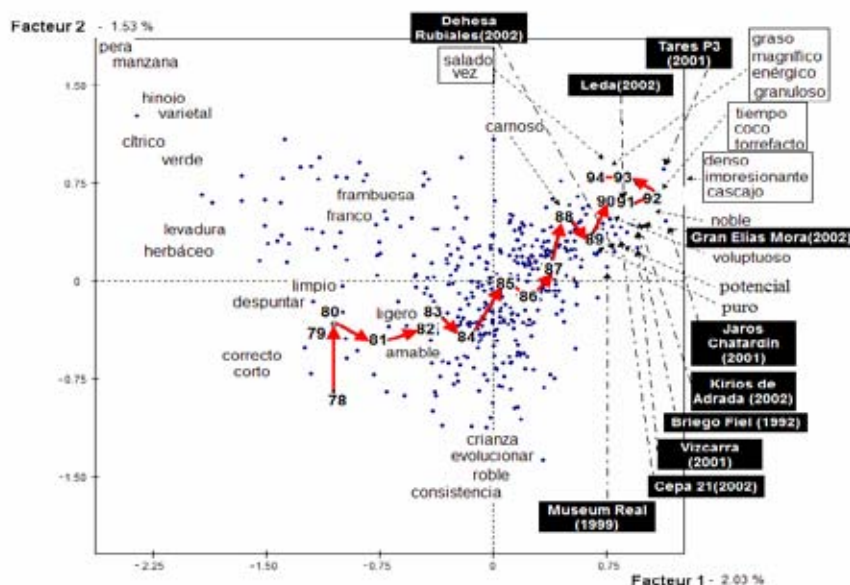


Figura 9. Primer plano factorial proporcionado por el AC de la tabla Vinos×Palabras

Posicionamiento de los diferentes valores de la nota considerados como categorías suplementarias

Otra manera de relacionar la tabla léxica y la puntuación consiste en efectuar el AC de la tabla léxica para después posicionar la puntuación como suplementaria. La posición de dicha puntuación sobre un eje corresponde a su correlación con el eje. Se observa una alta correlación entre la nota y el primer eje de AC (.70), lo que significa dicho eje expresa la nota en gran parte. El primer eje como un eje de calidad, aunque pueda existir una dirección de dispersión en la nube de vinos inducidas por tabla léxica más relacionada con la puntuación. La siguiente sección proporciona una herramienta para identificar dicha dirección.

4.2.4. Dar un papel simétrico activo a las palabras y a la puntuación, considerados como dos maneras de evaluar los vinos

Se construye la tabla múltiple, yuxtaponiendo la tabla léxica (con J columnas-palabra) y una columna con la puntuación dada a los vinos. Se analiza dicha tabla mediante el AFM, tratando la tabla léxica como en AFMTC. Este AFM extendido es equivalente a realizar un ACP ponderado:

- Sobre la tabla yuxtaponiendo la tabla, construida a partir de la tabla léxica, de término general $(f_j - f_{i \cdot} f_{\cdot j}) / (f_{i \cdot} f_{\cdot j})$ y la columna cuantitativa (centrada o centrada y estandarizada).
- Dando a las filas el peso inducido por el AC (es decir, $\{f_i; i = 1, \dots, I\}$);

- Dando a las columnas-palabra los pesos inducidos por el AC (es decir, $\{f_j; j=1, \dots, J\}$) divididos por el primer valor propio obtenido en el AC separado de la tabla léxica – notado γ_1^J – y a la columna cuantitativa un peso igual a 1.

Este AFM extendido proporciona los resultados clásicos del ACP, así como herramientas para comparar las diferentes estructuras inducidas sobre los vinos por los comentarios, por una parte, y la puntuación, por otra parte.

La figura 10 presenta el primer plano factorial inducido por este AFM. Se posicionan los diferentes valores de la puntuación como categorías suplementarias, es decir, en el centro de gravedad de los vinos que han puntuados con este valor.

El primer eje está muy fuertemente correlacionado con la nota (0.95). Como era de esperar, pocas palabras son francamente negativas: una guía de vinos sólo menciona vinos recomendables. Los aspectos hedónicos del vino se subrayan (*impresionante, espléndido*, a la derecha del eje, opuestos a *agradable, amable, medio*, a la izquierda). Lo que se suele llamar “la potencia del vino” parece ser un criterio dominante en la evaluación: los adjetivos *denso, graso, concentrado y largo* (en boca), a la derecha del eje, se oponen a *ligero o corto* a la izquierda. Algunos términos conciernen defectos (*sequedad, sobremaduro, evolucionar*); se encuentran en la parte izquierda del primer eje, indicando que cualifican los vinos peor puntuados.

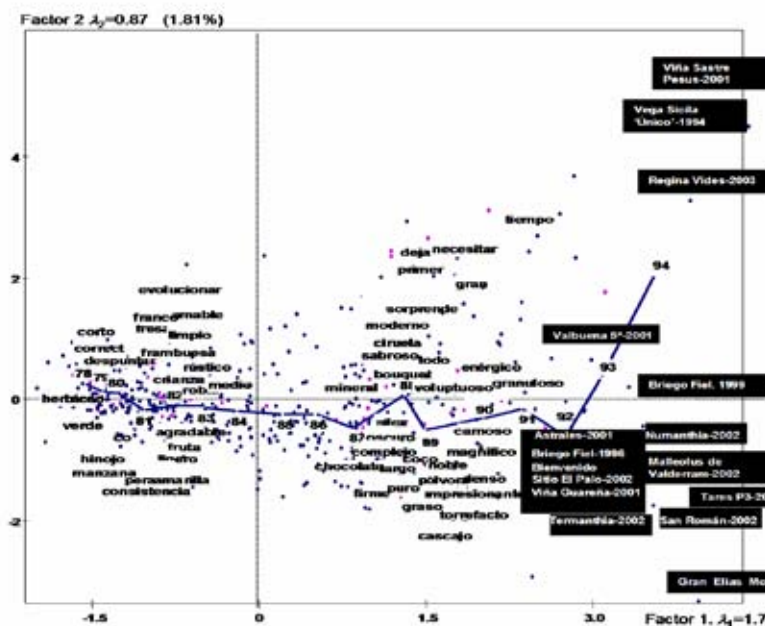


Figura 10. Primer plano factorial proporcionado por el AFM aplicada a la tabla yuxtaponiendo la tabla léxica y la columna con la puntuación

Algunas palabras – como *fácil, agradable y franco*–, consideradas palabras positivas en el lenguaje usual, no lo son tanto aquí, como lo indica su posición en la parte izquierda del eje. No es que los autores de la guía piensen que un vino excelente debe ser *desagradable, difícil y no franco*, sino

que oponen dichos términos a *complejo* o *sorprendente*. Apreciar el vino es también disfrutar de un punto de vista intelectual.

La trayectoria de la puntuación es bastante regular a lo largo del primer eje. El ordenamiento de las categorías no está alterado: el vocabulario presenta realmente una dimensión valorativa, bien expresada por este eje. Una validación de tipo bootstrap, no presentada aquí, ha permitido validar este hecho.

Segundo eje

El segundo y siguientes ejes presentan correlaciones bajas o nulas con la puntuación de los vinos. Además, no están correlacionados, o muy poco, con las otras características como la añada o el precio, estas últimas variables utilizadas como suplementarias en el análisis. Expresan la parte de la variabilidad del vocabulario no relacionado con dichas variables.

Los vinos con coordenadas positivas (respectivamente, negativas) sobre el segundo eje tienden a presentar comentarios más largos (respectivamente, más cortos) que lo esperado debido a su puntuación.

En el caso de los vinos mejor puntuados (parte positiva del primer eje), el segundo eje opone los vinos cualificados como *complejos*, *voluptuosos*, *sorprendentes*, etc. –que sugieren comentarios más largos– a los vinos *potentes*, *impresionantes*, *torrefactos*, *densos*, etc., más clásicos.

En el caso de los vinos con coordenadas negativas sobre el primer eje – por tanto, con puntuaciones bajas – el segundo eje destaca tres vinos considerados muy particulares (*Mesoneros de Castilla-2003*, *Valdelosfriales-2003* y *Fuentenarro-2002*). Dichos vinos son *jóvenes*, *típicos* y con pobres características *en boca*.

Discordancias entre los comentarios libres y la puntuación del vino

Aunque exista una muy fuerte correlación entre los comentarios libres y la puntuación, algunos vinos muestran comentarios más laudatorios (respectivamente, menos laudatorios) que su puntuación dejaba esperar. Esta discordancia está puesta de relieve mediante la representación superpuesta –no reproducida aquí– de las dos estructuras parciales. Por ejemplo, el punto parcial “comentarios libres” de *Tares P3-2001 premium* obtiene una coordenada mucho más positiva sobre el primer eje que el punto parcial “puntuación” correspondiente al mismo vino. De hecho, los comentarios de este vino contienen 8 de las 20 palabras de mayor coordenada sobre el primer eje (*impresionante*, *grande*, *a la vez*, *graso*, *cascajo*, *pólvora*, *largo* y *impresión*), una de ellas utilizada dos veces (*pólvora*).

5. Un experimento en curso: *parlem de vins en català i en francès!*

La interpretación de los datos de cata presentados en la sección 3 sufría de la ausencia de información suplementaria. No obstante, el estudio mostró que la descripción libre de los vinos comportaba una dimensión valorativa importante que conviene aprovechar.

Por esta razón, se decidió:

- Poner a prueba la contribución de las descripciones libres en sensometría, es decir, en la recogida y análisis de datos sensoriales mediante métodos estadísticos.
- Comparar las evaluaciones dadas por degustadores catalanes y franceses a un mismo conjunto de vinos.

Esta experiencia se realizó con éxito en la *Escola d'Enologia* (Espiells) y en el *Conseil Interprofessionnel des Vins du Roussillon-CIVR* (Perpiñan).

Cada sesión comportaba 2 ejercicios de cata distintos sobre los mismos 8 vinos catalanes:

- En la primera parte, cada degustador debía formar “clases de vinos” según la similitud percibida entre ellas. Los criterios de similitud eran los propios, sin que puedan existir buenas o malas respuestas. Después, las clases se apuntaban en la hoja proporcionada al degustador. Además, se pedía a cada uno que asocie a cada grupo las palabras que caracterizaban bien al conjunto del grupo.
- En la segunda parte, los mismos degustadores rellenaban una ficha de cata, en la cual debían describir los aspectos visual, olfativo y gustativo de cada vino; así rellenaban una ficha simplificada en comparación con la ficha empleada en la cata de 2003.

Además, se dispone de las características químicas de los vinos.

Los datos están capturados y en fase de depuración. La metodología presentada en la sección 4 permitirá analizar estos datos complejos, incluso las descripciones libres escritas o en catalán o en francés sin proceder a ninguna traducción. Las proximidades entre palabras inducidas por el análisis podrán sugerir equivalencias entre palabras.

6. Conclusiones

La codificación de los textos respeta tres puntos esenciales: estabilidad, legibilidad y reproducibilidad. Dicha codificación, relativamente fácil de implementar gracias a las herramientas elaboradas en la industria del lenguaje natural, asegura una codificación independiente del codificador y facilita la comparación entre estudios.

La metodología presentada permite analizar simultáneamente comentarios libres y variables cuantitativas y/o categóricas (Bécue & Pagès, 2008) y constituye una herramienta útil de numerosas áreas. En general:

- Las preguntas cerradas/ las puntuaciones aportan un marco sólido, pero tienden a reproducir las asunciones que han guiado la concepción del cuestionario cuando se analizan solas;
- Las preguntas abiertas/ los comentarios libres proporcionan información rica pero que difícilmente se puede analizar por sí sola.

Tonar en cuenta los dos tipos de cuestionamiento proporciona una herramienta que adiciona sus respectivas ventajas a la vez que limita los defectos propios de cada uno. Es obvio que el procedimiento presentado no permite capturar toda la información contenida en los comentarios o respuestas libres que constituyen un material muy complejo. No obstante, esta pérdida de información aporta a cambio una ganancia en comprensión, lo que es un principio básico de la estadística.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Education and Science, FEDER (grant SEJ2005-00741/ECON) as well as the Catalan Commission for Universities DURSI (grant SGR 00004/2005).

Agradecemos a Joan Miquel Canal, director del Parc Científic de les Indústries Enològiques de Falset habernos facilitado los datos de la cata del Priorat de 2003.

7. References

- Bécue-Bertaut, M., Pagès, J., (1999). Intra-sets multiple factor analysis. Application to textual data. In: Bacelar-Nicolau, H., Costa Nicolau, F., Janssen, J. (eds.), *Applied Stochastic Models and Data Analysis*. INE, Lisbon, 72-79.
- Bécue-Bertaut, M., Pagès, J., (2004). A principal axes method for comparing contingency tables: MFACT. *Comput. Statist. Data Anal.* 45(3), 481-503.
- Bécue-Bertaut, M., Pagès, J., (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Comput. Statist. Data Anal* 52(6) 3255-3268.
- Benzécri, J.P. (1973). *Analyse des Données, Vol. 2: Analyse des correspondances*. Paris: Dunod.
- Benzécri, J.P. (1981). *Pratique de l'analyse des données, Vol. 3, Linguistique & Lexicologie*. Paris: Dunod.
- Benzécri, J.P., (1983). Analyse de l'inertie intraclass par l'analyse d'un tableau de contingence. *Les Cahiers de l'Analyse des Données* 8(3), 351-358.
- El Mundo (2005). *Guía de catas 2005. Vinos de Castilla y León*. Valladolid: Biblioteca La Posada.
- Cazes, P., Moreau, J., (1991). Analysis of a contingency table in which the rows and the columns have a graph structure. In: Diday, E., Lechevallier, Y. (Eds), *Symbolic-numeric data analysis and learning*, Nova Science Publishers, New York, 271-280.
- Cazes P., Moreau, J., (2000). Analyse des correspondances d'un tableau de contingence dont les lignes et les colonnes sont munies d'une structure de graphe bistochastique. In: Moreau, J., Doudin, P.A., Cazes, P. (Eds), *L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données*, Springer, Berlin-Heidelberg, 87-103.
- Escayola J. (2008). Sensometría. Aportación a las descripciones libres. Proyecto de fi de carrera de la Diplomatura de Estadística de la UPC.
- Escofier, B., 1983. Analyse de la différence entre deux mesures sur le produit de deux mêmes ensembles. *Les Cahiers de l'Analyse des Données* 8, 325-329.
- Escofier, B. (2003). *Analyse des correspondances*. Rennes: Presses Universitaires de Rennes.
- Escofier, B., Drouet, D. (1983) Analyse des différences entre plusieurs tableaux de fréquence. *Les Cahiers de l'Analyse des Données* 8(4), 491-499.
- Escofier, B., & Pagès, J. (1988-1998) *Analyses factorielles simples et multiples*. Paris: Dunod.
- Labbé, D. (1990) *Normes de saisie et de dépouillement des textes politiques*. CERAT Cahier n° 7. <http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeNormes.pdf>

- Lebart, L. (2003) Analyse des données textuelles. In G. Govaert, *Analyse des données* (pp. 151-168). Paris: Lavoisier.
- Lebart, L., Piron, M., & Morineau, A. (2006) *Statistique exploratoire multidimensionnelle. Visualisation et inférence en fouille de données*. Paris: Dunod.
- Lebart, L., Salem, A., Bécue M. (2000). *Análisis estadístico de textos*. Milenio.
- Lehrer, A. (1975) Talking about Vino. *Language*, 51(4), 901-923.
- Muller, Ch. (1977-1992) *Principes et méthodes de statistique lexicale*, Paris: Larousse; reprint Genève: Champion-Slatkine.
- Murtagh, F. (2005). *Correspondence Analysis and Data Coding with R and Java*. Chapman & Hall, CRC Press, London
- Murtagh, F., Ganz, A., McKie, S. (In press) The structure of narrative: The case of film scripts. *Pattern Recognition* doi: 10.1016/J.patcog.2008.05.026.
- Orlaci (1975). *Multivariate analysis in vegetation research*. Junk. La Hague.
- Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval*. McGraw-New York: Hill Book Company.
- ten Kleij, F., & Musters, P.A.D. (2003) Text analysis of open-ended survey responses: a complementary method to preference mapping. *Food quality and preference*, 14, 43-52.

Cloenda

Bernat Joan

Secretari de Política Lingüística de la Generalitat de Catalunya

Bon dia, i moltes gràcies per haver-me convidat a participar en aquestes Primeres Jornades de Llengua i Estadística. En primer lloc, voldria remarcar la interdisciplinarietat d'aquestes jornades: el fet de posar en contacte persones que treballen en camps diferents permet compartir punts de vista, debatre, dialogar i, naturalment, obrir camins. Esperem que aquestes jornades tinguin continuïtat en el futur.

Un cop dit això, anem als continguts d'aquestes dies. Durant aquests dos dies les Jornades de Llengua i Estadística han tractat tres blocs de contingut.

D'una banda s'ha parlat de la qualitat de la llengua en la ciència estadística, de la terminologia, de la precisió en el llenguatge. També de l'adequació lingüística en la difusió de les activitats estadístiques en els mitjans de comunicació, és a dir, de la transparència del llenguatge usat. I finalment de la cura necessària en l'elaboració dels qüestionaris, principal instrument de recollida d'informació d'una enquesta. En resum, el que s'ha tractat en aquestes Jornades de Llengua i Estadística ha estat la llengua com a instrument de la ciència estadística.

Però les Jornades de Llengua i Estadística podrien haver tractat també de l'estadística com a instrument de la llengua o de la política lingüística. I això, si em permeteu, és el que m'agradaria comentar-vos durant uns minuts: les estadístiques de la llengua.

Des dels inicis la política lingüística que ha dut a terme el Govern de la Generalitat ha tingut com a base l'estadística. Des de l'any 1981, amb el primer padró, que contenia preguntes sobre coneixements de llengua catalana, fins a l'any 2008, amb la realització de la segona edició de l'*Estadística d'usos lingüístics de la població*, les decisions s'han pres a partir de l'anàlisi, la interpretació i la representació d'aquestes dades.

Volem remarcar, però, el salt qualitatiu que hi ha entre les dades de coneixements lingüístics que recullen censos i padrons i les dades que ens ofereix l'*Estadística d'usos lingüístics*. L'avenç ha estat importantíssim. Si fins aleshores disposàvem de dades sobre coneixements de la llengua catalana, amb aquesta estadística disposem de dades empíriques sobre la identificació lingüística de la població i sobre els usos que aquesta població fa del català i de les altres llengües parlades a Catalunya en els diversos espais de relació personal.

Certament no és un cens, però es tracta d'una enquesta a més de 7.000 persones, amb un qüestionari de 141 preguntes que recullen informació no només sobre els coneixements de català sinó també sobre la llengua inicial, la llengua d'identificació i la llengua habitual, els usos lingüístics a la llar, a la família, en els àmbits socials i com a persones consumidores o usuàries de serveis.

A més, en aquesta segona edició de l'estadística per primer cop hem inclòs preguntes sobre coneixements de castellà (i també anglès i francès) i hem establert una mostra de la població de la Vall d'Aran suficient per poder tenir dades estadísticament representatives dels usos lingüístics de l'occità, tercera llengua oficial de Catalunya.

El Pla estadístic de Catalunya inclou, a més de l'*Estadística d'usos lingüístics de la població* de què us he parlat, l'*Estadística d'usos lingüístics en institucions i empreses*, que ens permet disposar d'una descripció de la llengua en el sector econòmic públic i privat.

D'altra banda, cada cop més enquestes inclouen preguntes sobre la llengua de l'entrevistat. Aquestes preguntes, a banda de donar molta informació addicional a l'objecte de l'enquesta, enriqueixen la investigació si es creuen amb altres variables de l'enquesta com es fa amb la resta de variables sociodemogràfiques. A més, una simple variable d'observació com és la llengua en què es respon el qüestionari, en països amb més d'una llengua, ajuda a l'anàlisi de la informació obtinguda i és fonamental per assegurar la qualitat del procés d'enquestament.

Cal remarcar que aquestes qüestions no només es plantegen en països amb més d'una llengua com Catalunya, sinó en molts altres països occidentals, oficialment monolingües però que reben grans contingents de població immigrada que estan transformant la seva realitat social i lingüística.

Totes aquestes reflexions són adequades perquè l'ordenació de les llengües en l'espai públic esdevé avui dia una necessitat en molts països, generalitzada cada cop més, independentment de l'oficialitat lingüística. Així, la política lingüística, entesa com a política social i pública, necessita la credibilitat de l'estadística oficial.

Jornades com aquesta afermen aquesta credibilitat, perquè cerquen, en la reflexió i l'anàlisi, el rigor.

No em queda més que felicitar-vos per la iniciativa que representen aquestes Jornades, esperar que tinguin continuïtat en el futur i sobretot pensar que s'ha començat a obrir un camí que convergeix amb la tasca que s'ha realitzat ja durant molts anys, amb la innovació tant en estadística com en anàlisi de la llengua.

Moltes gràcies.

