Treballs de Sociolingüística Catalana, núm. 26 (2016), p. 181-197 DOI: 10.2436/20.2504.01.115
http://revistes.iec.cat/index.php/TSC ISSN (ed. impresa): 0211-0784 ISSN (ed. electrònica): 2013-9136

Predicció de l'ús del català mitjançant la classificació supervisada

Predicting the use of Catalan through supervised classification

Francisco Grimaldo,* Emilia López-Iñesta,* Manel Perucho* i Ernest Querol**[1]
*Universitat de València
**Universitat Oberta de Catalunya

Data de recepció: 25 de març de 2015
Data d'acceptació: 16 de juny de 2015

Resum

Un dels principals reptes que ha tingut i té la sociologia del llenguatge és esbrinar quines són les variables que influeixen en els usos lingüístics. En la recerca que presentem ens valem dels mètodes d'una àrea de la intel·ligència artificial, l'aprenentatge automàtic (*machine learning*), que estudia la implementació de mètodes computacionals que permeten induir models de coneixement a partir d'informació que prové de dades d'exemple disponibles, per a escatir si algun d'aquests millora la predicció del grau d'utilització de la llengua catalana aconseguida fins ara. Hi hem fet servir tres mètodes de classificació supervisada: Naive Bayes, arbres de decisió i màquines de vectors de suport.

Per a complir aquesta comesa calia un corpus empíric que ens permetera tant la comprovació del nivell de predicció d'un model teòric com la seua validesa en diferents contextos sociolingüístics. Les recerques que coneixem que tenen uns percentatges més alts de predicció són les dutes a terme per Querol, que han estat avaluades en tots els territoris on es parla català. La investigació que hem fet amb aquestes dades permet concloure que la classificació supervisada pot servir per a construir models de predicció del grau d'ús del català amb un percentatge d'encert que supera els aconseguits en les investigacions precedents. Amb la qual cosa podem establir quines són les variables més informatives. A més, també ens ajuda a resoldre el problema metodològic de la divisió en grups lingüístics i palesa que l'ús és un sistema continu.

Paraules clau: ús lingüístic, predicció, intel·ligència artificial, aprenentatge automàtic, classificació supervisada.

Correspondència: Francisco Grimaldo. Universitat de València. Departament d'Informàtica. Avinguda de la Universitat, s/n. 46100 Burjassot. A/e: *francisco.grimaldo@uv.es*. A/I: *http://www.uv.es/grimo*. Tel.: 963 544 487. Fax: 963 544 768.

Correspondència: Emilia López-Iñesta. Universitat de València. Departament d'Informàtica. Avinguda de la Universitat, s/n. 46100 Burjassot. A/e: *eloi@alumni.uv.es*.

Correspondència: Manel Perucho. Universitat de València. Departament d'Astronomia i Astrofísica. Avinguda de Vicent Andrés Estellés, s/n. 46100 Burjassot. / Universitat de València. Observatori Astronòmic. Carrer del catedràtic José Beltrán, 2. 46980 Paterna. A/e: *manel.perucho@uv.es*.

Correspondència: Ernest Querol. Universitat Oberta de Catalunya. Departament d'Arts i Humanitats. Avinguda del Tibidabo, 39-43. 08035 Barcelona. A/e: *equerolp@uoc.edu*.

1. Tots els autors formen part del grup de recerca Alcàntera: el Pont amb la Sociolingüística.

Abstract

One of the main challenges that the sociology of language has faced is the determination of the variables that govern the use of a language. Inspired by the field of artificial intelligence, in this study we make use of machine learning as a suitable approach to implement computational methods that permit the induction of linguistic use models derived from the available data. We aim to improve the level of prediction for the degree of use of the Catalan language achieved up to now. To this end, we have used three supervised classification techniques: Naive Bayes, decision trees, and support vector machines.

We needed an empirical corpus that would allow us to test the prediction level of a theoretical model as well as its validity within different sociolinguistic situations. To the best of our knowledge, the work by Querol is the one providing the highest prediction success in all the Catalan-speaking territories. Thus, the research presented in this paper uses that data to conclude that supervised classification can be used to successfully determine prediction models for the degree of use of Catalan that outperform previous attempts and that allow us to identify the most relevant variables of the problem. Moreover, it also helps us to solve the methodological problem of the division of linguistic groups and shows that the use of a language is a continuous system rather than a discrete one.

Keywords: linguistic use, prediction, artificial intelligence, machine learning, supervised classification.

1. Introducció

La sociologia del llenguatge s'ocupa fonamentalment de l'anàlisi dels usos lingüístics i d'establir quins factors es relacionen amb aquests. En la investigació que ara presentem pretenem esbrinar si alguna tècnica de classificació supervisada pot millorar la predicció del grau d'utilització de la llengua catalana aconseguida fins aquest moment.

Per a incrementar els percentatges de predicció de l'ús lingüístic ens cal analitzar les dades de les recerques que hagen obtingut els resultats més alts (Querol, 1999). Són justament les que s'allunyen de l'esquema unicausal d'explicació que es pretenia obtenir amb diferents teories clàssiques (les actituds lingüístiques, la motivació, la identitat, la vitalitat etnolingüística, etc.). Considerem que el fet de proposar la interacció de tres variables per a explicar l'ús és una de les raons per les quals el seu model teòric aconsegueix el nivell de predicció més alt. A més, el fet que haja passat enquestes en tots els territoris on es parla català ens permetrà analitzar els resultats en territoris amb contextos diferents.

2. Corpus empíric i teòric

El present article enllaça, doncs, amb tota la recerca feta des del paradigma sociològic de la definició social per a l'anàlisi dels processos de substitució i de reversió

lingüística. Aquesta aproximació, proposada per Ernest Querol (1999, 2000, 2002*a* i 2002*b*) i inscrita dins del marc de la teoria de les representacions socials, va significar un nou enfocament teòric basat en una analogia amb la teoria matemàtica de les catàstrofes. La investigació empírica que ha fet aquest autor ha estat el fruit de tot aquest plantejament teòric que intentava millorar els resultats obtinguts mitjançant les teories clàssiques i s'adreçava a respondre una llarga sèrie de preguntes, la principal de les quals era: quines variables influeixen en els usos de les llengües?

El contrast entre el nombre de recerques dutes a terme seguint la teoria de les actituds lingüístiques i els fruits obtinguts (que rarament havien depassat correlacions del 30 % entre aquestes i l'ús de les llengües) és el que li va fer plantejar la necessitat d'un canvi de marc teòric i metodològic que, al capdavall, és el que proposen les seues recerques. L'autor partia del convenciment que l'esquema unicausal (açò és, només una variable és la que explica l'ús de les llengües) era massa estret per a donar compte de la complexitat de les tries lingüístiques i, en definitiva, per a donar-ne una explicació. Li calia, per tant, analitzar sobre quins fonaments s'havien construït les teories explicatives i aquesta recerca faria necessari el plantejament i la resolució de diferents qüestions. Per a subratllar-ne les fonamentals, en primer lloc li va caler fer servir, per primera vegada en la sociolingüística catalana, el concepte de *representació social*, que prové de la psicologia social (Jodelet, 1989). En segon lloc, el desllorigador per a decidir el nombre de variables que havien d'intervenir en el model i quines havien de ser li'l va aportar el recurs a la teoria de les catàstrofes. L'analogia amb aquesta teoria (que va arribar fins a la relació biunívoca) li va servir per a dur a terme una reflexió discursiva d'acord no tan sols amb les normes empíriques sinó també amb un mínim sistema formal.

Passem, sense més preàmbuls, a explicar aquest model. Com ja hem avançat i és comú a moltes recerques en aquesta disciplina, l'ús de les llengües era clarament la variable dependent; per tant, les recerques en els diferents territoris on es parla català es van estructurar a partir de la hipòtesi general següent: els tipus de comportaments lingüístics[2] són el resultat de la interacció d'un conjunt de representacions socials que es fa el subjecte. Com mostra el gràfic 1, hi té particular importància la interacció entre les tres següents: la representació de cadascuna de les llengües en presència, la representació de la xarxa interpersonal de comunicació[3] i la representació del grup social de referència.[4]

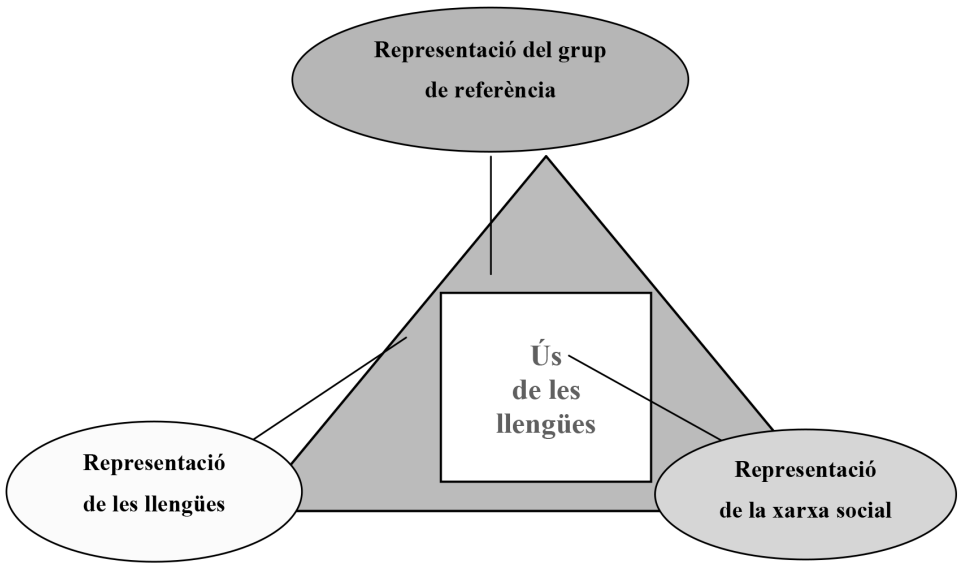
2. Es considera que les possibilitats generals de comportaments lingüístics en la situació de relació interlingüística més simple (la de contacte entre dues llengües L1 i L2, tot i que es proposa que per a una anàlisi més acurada cal tenir sempre present la tríada) són: monolingüisme en L1, bilingüisme +L1 que L2, bilingüisme L1/L2, bilingüisme +L2 que L1 i monolingüisme L2.

3. Félix Requena (1989: 137) assenyala que s'ha de concebre el concepte de *xarxa social* no pas de manera metafòrica sinó des d'un punt de vista analític. Per a la recerca sociolingüística entenem la xarxa social com el conjunt de persones que comparteixen una de les seues llengües i que tenen interaccions lingüístiques entre elles.

4. Merton i Kitt (1950: 101-102) caracteritzen el grup de referència com segueix: «Apunta a sistematitzar els determinants i les conseqüències dels processos d'avaluació i d'autoavaluació en els quals

Les recerques empíriques fetes en tots els llocs on es parla català van demostrar la validesa d'aquesta hipòtesi general (i de les hipòtesis secundàries que es van desenvolupar a partir d'aquella) i van permetre corroborar el marc teòric i metodològic en què se sustenten amb uns resultats que gairebé van triplicar el percentatge d'encert respecte als models anteriors (Calaforra, 2002; Kremnitz, 2002). Vegem-ho en l'epígraf següent.

GRÀFIC 1
Model explicatiu dels usos lingüístics (variable dependent) i la interrelació amb les variables independents



FONT: Querol (2000: 90).

2.1. *Predicció del grau d'ús del català aconseguida amb anterioritat*

El treball empíric dut a terme per a validar les hipòtesis anteriors es va emmarcar en un projecte d'obtenció de dades de tots els territoris on es parla català: Catalunya (1993-2000-2004), el País Valencià (1998), les Illes Balears (2001), Andorra (2002), la Franja d'Aragó (2004), la Catalunya del Nord (2004) i l'Alguer (2004). La taula 1 mostra els detalls de la fitxa tècnica per a la recollida de dades.

---

l'individu assumeix els valors o les normes d'altres individus i grups com un marc de referència». Boudon *et al.* (1993: 109) el defineixen com a «grup les actituds, comportament, creences o valors del qual són adoptats com a criteri per un individu quan defineix una situació, la valora o decideix actuar. El grup de referència pot ser un grup, un individu o fins i tot una idea i té una doble funció, comparativa i normativa».

TAULA 1
Fitxa tècnica d'obtenció de dades

Àmbit territorial	Cada territori sencer estudiat.
Unitat d'anàlisi	S'ha realitzat un mostreig en etapes múltiples. En la primera etapa es van escollir les unitats primàries, que són els instituts d'ensenyament secundari. En la segona, es van triar dins de cada institut les unitats secundàries, que són els cursos de 4t d'ESO.[5] I, finalment, en la tercera etapa, es va seleccionar l'aula de cada institut en què s'havia de passar l'enquesta, que constitueix la tercera unitat d'anàlisi.
Univers estadístic	La totalitat d'estudiants de 4t d'ESO tant de centres públics com privats.
Disseny de la mostra	Aplicació de la tècnica de mostreig per conglomerats.
Efectius de la mostra i marge d'error	El nombre de persones enquestades ha estat sempre de més de 380, seleccionades segons el mostreig en etapes múltiples per conglomerats que hem exposat. L'error mostral ha estat del 5 % per a un nivell de confiança de 2 sigma.
Qüestionari	El qüestionari inclou 378 preguntes i sempre ha estat el mateix, llevat d'Andorra i de l'Alguer, atesa la situació de quadrilingüisme (català, castellà, portuguès i francès) a Andorra i de trilingüisme a l'Alguer (català, sard i italià).[6] Pel que fa a Catalunya, els tres anys s'han fet les mateixes preguntes en els mateixos instituts d'ensenyament secundari.
Treball de camp	Cada enquestador (un per llengua) parlava en una llengua i sempre es donaven les mateixes indicacions i en el mateix ordre de llengües, es començava parlant en català i l'altre enquestador continuava en castellà. Els estudiants podien triar els qüestionaris en la llengua que volien.

Aplicant el model teòric presentat es pot fer una predicció sincrònica que, a partir de les variables independents (açò és, la representació de les llengües, de la xarxa social i del grup de referència), prediga quins usos faran els estudiants que han respost les enquestes (Querol, 2004*a*, 2004*b* i 2005). El nivell d'encert inicial d'aquest model en l'enquesta de Catalunya per a l'any 1993 es va situar en un valor mitjà del 83,8 % per als quatre comportaments lingüístics considerats. Tanmateix, aquest valor descendia al 72,3 % en l'enquesta de Catalunya de l'any 2000 i encara baixava un poc més en les recerques d'Andorra (71,8 %), el País Valencià (64,4 %) i les Illes Balears (60,1 %).

5. A Catalunya es va passar a segon de BUP, que va ser el curs escollit el 1993, però, atesa la implantació del nou sistema educatiu, el 2000 i el 2004 es va decidir que també s'havia de passar a un curs equivalent de l'ensenyament postobligatori, com ho era segon de BUP. Així, es va triar el primer de batxillerat. A la resta de territoris es va adoptar el criteri que fóra el darrer any de la secundària obligatòria.

6. Només s'han fet adaptacions a les varietats lingüístiques de cada territori. A les comunitats autònomes de l'Estat espanyol s'ha passat en català i en castellà; a Andorra, només en català, i a la Catalunya del Nord i a l'Alguer, només en francès i només en italià, respectivament.

Com ja avançàvem en la introducció, en aquest article emprarem un conjunt de tècniques de classificació supervisada per a comprovar si podem augmentar aquests nivells d'encert assolits fins ara per a predir el grau d'ús del català. En concret, farem servir els classificadors bayesians, els arbres de decisió i les màquines de vectors de suport.

3. Mètodes: la classificació supervisada

L'aprenentatge automàtic (també conegut pel terme anglosaxó *machine learning*) és una àrea de la intel·ligència artificial que estudia com implementar mètodes computacionals que permeten induir models de coneixement a partir d'informació que prové de dades d'exemple disponibles (Mitchell, 1997). En les últimes dècades, l'aprenentatge automàtic ha esdevingut una eina molt important en l'automatització de la presa de decisions i presenta múltiples aplicacions en la vida real, com ara la salut, en què permet fer diagnòstics de malalties a partir de dades mèdiques; la recuperació d'informació, on resulta útil en la cerca de documents, fotografies o vídeos similars a un exemplar donat, o, fins i tot, la seguretat i la televigilància, a través del reconeixement facial de persones.

Un dels principals camps de l'aprenentatge automàtic és la classificació supervisada, en la qual diferents algorismes tracten d'aprendre un model matemàtic entrenat a partir d'un conjunt d'exemples etiquetats (açò és, se'n coneix la categoria o classe a què pertany cadascun) per tal de realitzar amb èxit, posteriorment, la predicció de les classes dels nous exemples no categoritzats. Així, es pot entendre la classificació com un mètode de generalització, en què a partir d'un conjunt inicial de dades d'entrenament es dedueixen unes característiques pròpies dels exemples de cada classe (denominades en la bibliografia *atributs*) amb les quals el model aprèn a classificar, proveït tant de la capacitat de distingir exemples entre si com la d'agrupar-los en classes.

Hi ha múltiples tècniques de classificació (o classificadors); per exemple: els arbres de decisió, les xarxes neuronals, els algorismes genètics, les regles d'associació, etc. L'aplicació d'una tècnica o d'una altra depèn de les característiques de la qüestió a resoldre, ja que cadascuna d'aquestes pot ser útil en unes circumstàncies determinades. En aquest treball experimentarem amb alguns algorismes de classificació representatius de diferents enfocaments clàssics. En particular, emprarem els classificadors bayesians, els arbres de decisió i les màquines de vectors de suport, ja que tots permeten aprendre models flexibles i explicatius capaços d'assignar una categoria de comportament lingüístic a un individu descrit pels tres índexs de representació social: la de la llengua, la del grup de referència i la de la xarxa social.

3.1. *Classificadors bayesians: Naive Bayes*

Els classificadors bayesians són classificadors estadístics basats en el teorema de Bayes que poden obtenir les probabilitats que un conjunt de dades (una mostra) pertanga a una classe particular o a una altra. Naive Bayes (NB) és el classificador bayesià més senzill i s'utilitza quan es vol classificar un exemple x descrit per un conjunt d'atributs a_j

dins d'un conjunt *C* que té un valor finit de classes c_i. D'acord amb el teorema de Bayes, aquesta classificació es pot expressar mitjançant la probabilitat *a posteriori* següent:

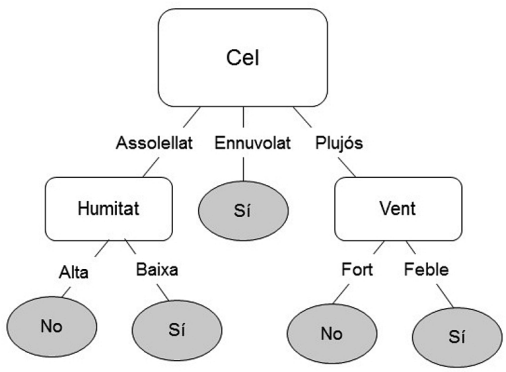
$$P(C = c_i|a_j) = \frac{P(a_j|c_i)\ P(c_i)}{P(a_j)}$$

El terme *naive* ('ingenu') fa referència al fet que el classificador pressuposa que els atributs de cada exemple són independents entre si, la qual cosa constitueix una premissa *forta* i poc freqüent en la realitat. No obstant això, el model sovint aconsegueix bons resultats (Mitchell, 1997), ja que la funció de classificació del mètode Naive Bayes classifica els nous exemples d'acord amb el valor més probable atenent els valors de les seues característiques.

3.2. *Arbres de decisió*

Els arbres de decisió són un dels mètodes d'aprenentatge supervisat més utilitzats en la bibliografia clàssica. Poden interpretar-se com un conjunt de condicions elaborades a partir de les dades disponibles per a l'aprenentatge i organitzades en una estructura jeràrquica en forma d'arbre que poden utilitzar-se per a tasques de classificació o regressió. Un arbre de decisió està compost per un node arrel, nodes interns de decisió, nodes fulla i branques. Com mostra l'exemple del gràfic 2, cada node representa un atribut de les dades d'entrenament i les branques corresponen als possibles valors d'aquest atribut, mentre que els nodes fulla contenen el valor de classe assignat.

GRÀFIC 2
Exemple d'arbre de decisió per al concepte jugar al tennis



FONT: Adaptació de Mitchell (1997).

Així, doncs, l'objectiu dels arbres de decisió és obtenir les regles o relacions que permeten determinar la classe d'un exemple mitjançant el recorregut de l'arbre segons els valors dels seus atributs. Convé fer palès que, segons els algorismes de construcció de l'arbre i de les mesures de selecció d'atributs per a cada node, es pot induir

més d'un arbre de decisió a partir del mateix conjunt de dades d'entrenament. En aquest article emprarem un arbre de classificació de tipus J48 amb un criteri de maximització d'entropia que triarà primer aquells atributs més discriminants a l'hora de determinar el comportament lingüístic dels individus.

3.3. *Màquines de vectors de suport*

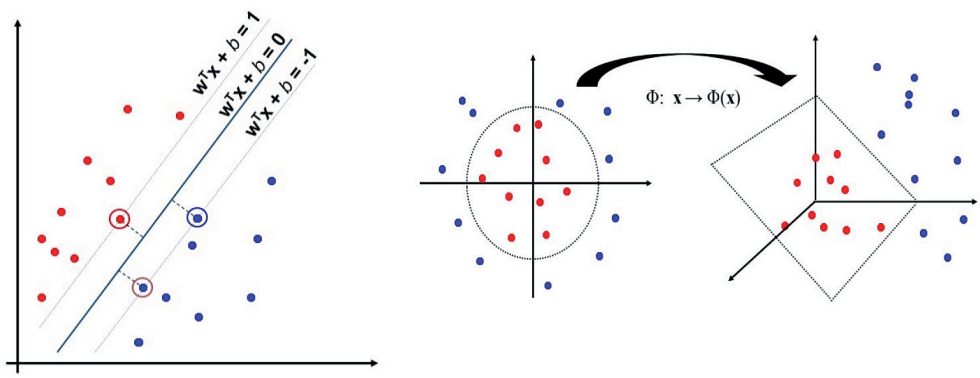
Les màquines de vectors de suport (en anglès, *support vector machines* o SVM), proposades originàriament per Boser, Guyon i Vapnik (1992), són una de les tècniques de classificació emprades més profusament avui. L'enfocament de les SVM es considera un bon candidat per a aquest treball a causa de la seua alta capacitat de generalització, fins i tot quan la dimensió de l'espai d'entrada és molt alta (Chapelle, Haffner i Vapnik, 1999).

En la seua vessant més bàsica, les SVM són classificadors binaris lineals que fonamentalment cerquen un separador lineal o hiperplà que discrimine entre dues classes segons una estratègia de maximització del marge. Aquesta tècnica consisteix a escollir, entre tots els hiperplans possibles, aquell que separe els punts pertanyents a dues classes diferents de manera que la distància als vectors més pròxims de cada classe siga màxima (aquests són els denominats *vectors de suport*, assenyalats amb un cercle en el gràfic 3*a*). Com que, en la pràctica, la majoria dels problemes no tenen una estructura lineal, la classificació mitjançant SVM pot ser transformada fàcilment a una classificació no lineal si es projecten els exemples del conjunt de dades a un espai de característiques de major dimensionalitat fent servir una transformació no lineal, de manera que el nou espai esdevinga linealment separable, tal com indica el gràfic 3*b*.

GRÀFIC 3

a) *Hiperplà de màxima separació i vectors de suport*

b) *Transformació d'un espai linealment no separable mitjançant una transformació no lineal*



FONT: *www.cs.utexas.edu/~mooney/cs391L/slides/svm.pdf.*

Encara que les SVM, en la seua formulació original, van ser dissenyades per a resoldre problemes de classificació binària, també es poden emprar en problemes de classificació multiclasse gràcies als mètodes «un contra tots» (Vapnik, 1995) o «un contra un» (Knerr, Personnaz i Dreyfus, 1990). L'experimentació que presenta aquest article fa servir un nucli radial com a funció de transformació i el mètode «un contra un» per a classificar els comportaments lingüístics en un context multiclasse no lineal.

3.4. *Disseny de l'experimentació*

Per a l'entrenament dels models de classificació introduïts suara, utilitzarem els resultats de les enquestes realitzades per Querol al llarg dels anys per a tots els territoris on es parla català en un context de convivència preeminent amb el castellà, açò és: Catalunya (1993-2000-2004), el País Valencià (1998), les Illes Balears (2001) i la Franja d'Aragó (2004).[7] La taula 2 resumeix les característiques dels índexs de representació social que emprarem en la construcció i avaluació dels models. Com a conjunt d'entrenament se seleccionarà aleatòriament un 70 % de les dades disponibles per a cada territori i dedicarem el 30 % restant al conjunt de test.

Taula 2
Índexs de representació social obtinguts de les enquestes

Índex	*Tipus*	*Interval*
Representació del català	Numèric	[86, 774]
Xarxa social en català	Numèric	[22, 198]
Grup de referència	Numèric[8]	{1, 2, 3, 4}
Grau d'ús del català	Numèric	[15, 135]

Malgrat que les dades originals representen el grau d'utilització del català com a variable numèrica, en aquest treball proposem la discretització de l'índex «grau d'ús del català» en un conjunt de categories *flexible* que represente els comportaments lingüístics que un individu pot adoptar. Per *flexible* entenem, d'una banda, que el nombre de categories (i, per tant, la grandària de cada subinterval) pot ser modificat mitjançant un paràmetre de l'experimentació (*NCategories*) per a ajustar el grau de

7. Resta com a treball futur l'extensió de l'experimentació a les zones singulars on el català conviu amb altres llengües, com ara la Catalunya del Nord, l'Alguer i Andorra.

8. En la bibliografia s'aconsella a vegades representar les variables categòriques com un conjunt de variables numèriques. Hem provat de desagregar el grup de referència per mitjà de dues variables que es corresponguen amb la representació binària de cada categoria i també hem provat de representar aquesta variable senzillament de forma categòrica. Tanmateix, els resultats obtinguts en ambdós casos han estat pitjors per a tots els classificadors que quan es considera el grup de referència directament com a variable numèrica.

precisió desitjat. D'una altra banda, per avaluar el nivell d'encert dels models de classificació construïts, mesurarem el percentatge d'individus del conjunt de test que han estat assignats a la seua categoria o a una categoria contigua. La definició d'aquestes fronteres flexibles s'ajustarà també mitjançant un paràmetre de l'experimentació que fa referència a la màxima distància a la qual es pot trobar una categoria per a considerar-la contigua (*DistContigua*).

A més d'entrenar els tres classificadors proposats (NB, J48 i SVM) amb les dades disponibles per a cada territori i any (aproximació que ens permetrà construir un model local per a un moment determinat en el temps), entrenarem el model que generalment proporciona millors prediccions (NB, com es veurà en l'apartat següent) amb la resta de dades provinents de les enquestes realitzades en altres territoris i anys. L'objectiu d'aquest darrer experiment serà avaluar la hipòtesi de si l'ús d'informació global ens ajuda a construir un model més ric que genere prediccions locals més encertades.

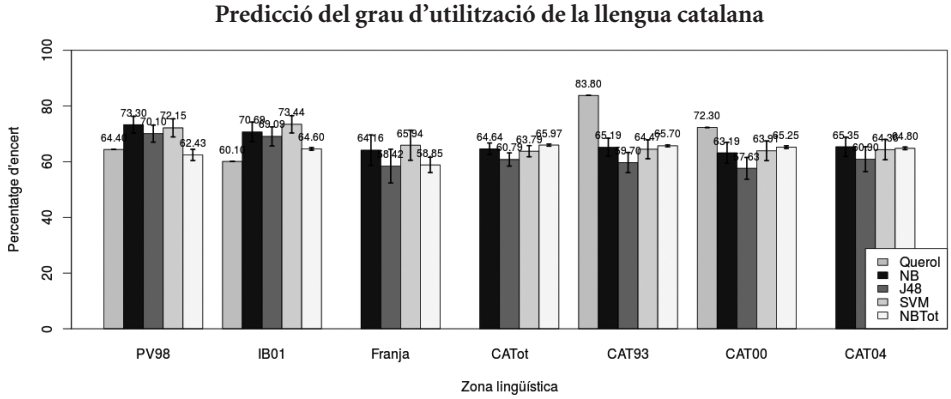
4. Resultats

Prenent com a punt de partida les prediccions de l'estat de la qüestió descrites en l'epígraf 2.1, el primer escenari en què hem avaluat els models de classificació supervisada proposats distingeix quatre comportaments lingüístics (*NCategories = 4*). Aquest escenari reprodueix un context sense *bilingües purs* (amb un ús igual de les dues llengües), on els grups presents són: exclusius castellanoparlants, més castellanoparlants, més catalanoparlants i exclusius catalanoparlants.

El gràfic 4 mostra el percentatge d'encert quan no acceptem les assignacions a grups contigus (*DistContigua = 0*). Tots els valors que obtenim es corresponen amb la mitjana i la desviació estàndard per a 100 repeticions de cada experiment. Observeu que les prediccions per al País Valencià (PV98) i per a les Illes Balears (IB01) milloren en aproximadament 10 punts percentuals els nivells d'encert assolits en treballs anteriors, mentre que per a Catalunya (CAT93 i CAT00) obtenim resultats més baixos en aquesta configuració, al voltant d'un 65 % d'encert. Per a la resta de territoris i anys, com ara la Franja d'Aragó (Franja) o les dades de Catalunya per a l'any 2004 (CAT04) i per a tots els anys (CATot), no hi ha dades possibles per a fer la comparació.

En general, els classificadors NB i SVM proporcionen millors prediccions, tot i que, dins dels marges d'error, els tres mètodes són equivalents. Aquest comportament és l'esperat, ja que la classificació probabilística que realitza NB així com la no lineal que realitza la SVM són més potents a l'hora d'enfrontar-se a un problema de subdivisió d'un espai de variables amb fronteres difuses (com les que separen els comportaments lingüístics d'una societat) que la divisió lineal que es deriva de qualsevol arbre de decisió. Ara bé, com veurem cap al final d'aquest epígraf, convé no descartar els arbres de classificació J48, ja que proporcionen una explicació intel·ligible dels models de predicció d'ús del català que interpretarem en clau sociopolítica. Cal destacar també el fet que el nivell d'encert siga molt semblant per a tots els territoris, la qual cosa palesa l'aplicabilitat dels models de classificació a qualsevol territori de l'àmbit lingüístic català.

GRÀFIC 4
Percentatge d'encert per a quatre categories de comportament lingüístic

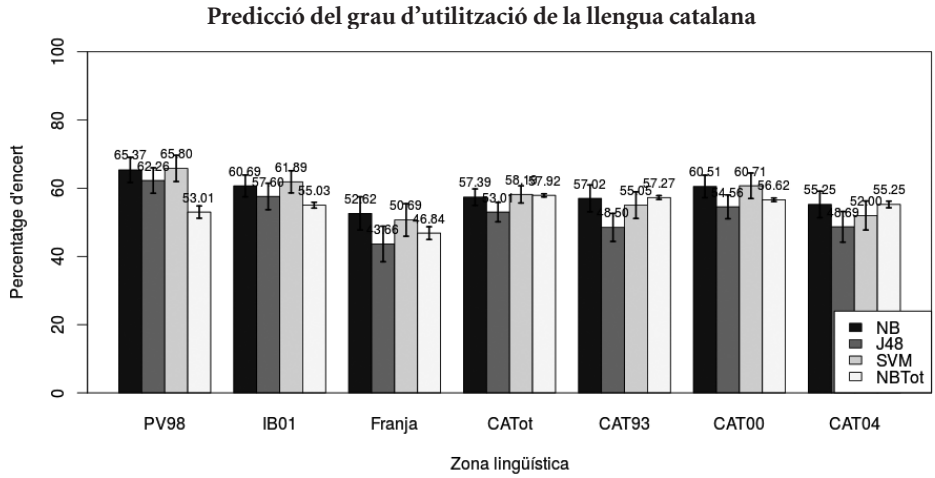


El gràfic 5 mostra el percentatge d'encert dels models de classificació quan afegim un grup de *bilingües purs* a l'escenari anterior, açò és, quan classifiquem d'acord amb cinc categories (*NCategories = 5*) i quan la precisió de l'encert continua essent màxima (*DistContigua = 0*). La disminució generalitzada del nivell d'encert per a tots els territoris corrobora la certesa que pretendre augmentar el grau de precisió desitjat requereix la definició de fronteres més difuses.

Per tant, plantegem la representació d'aquests cinc comportaments lingüístics mitjançant un desdoblament en deu categories (*NCategories = 10*) acompanyat de la definició d'una frontera flexible d'un grau de contigüitat (*DistContigua = 1*). Noteu en el gràfic 6 que aquesta aproximació fa augmentar el percentatge d'encert dels models de classificació en tots els territoris, especialment en el cas de Catalunya (CAT93, CAT00, CAT04 i CATot) i de la Franja. Els resultats ens permeten predir amb un grau d'exactitud superior al 70 % (que arriba a vegades a superar el 80 %) l'ús de la llengua catalana d'un individu coneixent els seus valors de la representació del català, de la xarxa social i del grup de referència.

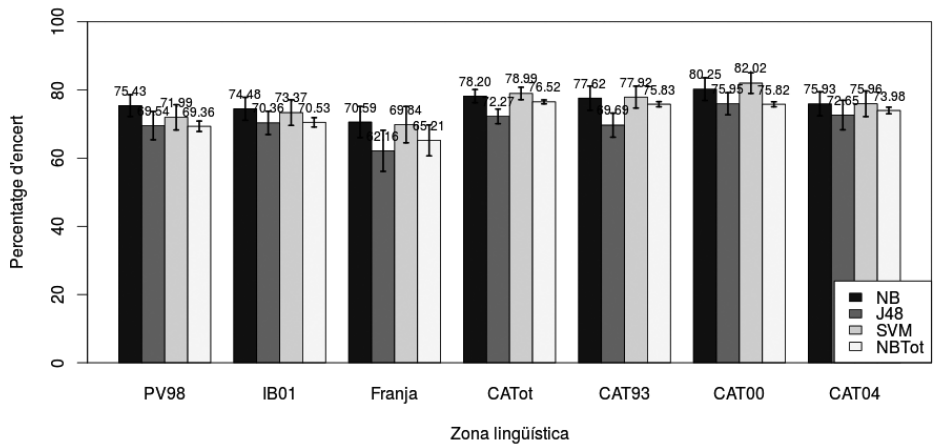
GRÀFIC 5

Percentatge d'encert per a cinc categories de comportament lingüístic



GRÀFIC 6

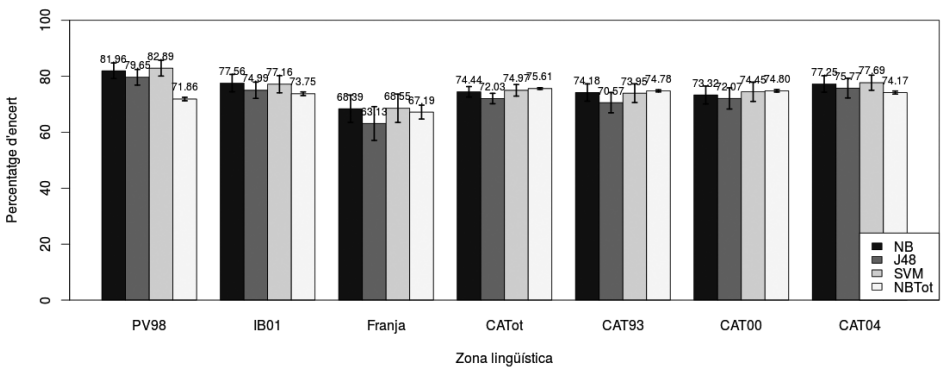
Percentatge d'encert per a deu categories flexibles de comportament lingüístic



Hom podria pensar que l'èxit de l'experimentació anterior, que compara quatre categories estrictes enfront de deu de flexibles, pot venir del fet que, en el primer cas, una categoria acceptable representa un 25 % de l'interval d'ús lingüístic, mentre que, en el segon cas, en representa un 30 %. Per estudiar aquest fet, hem comparat els resultats per a tres comportaments lingüístics (*NCategories = 3*) sense acceptar les assignacions a grups contigus (*DistContigua = 0*) i per a nou comportaments lingüístics (*NCategories = 9*) amb una frontera flexible d'un grau de contigüitat (*DistContigua = 1*). Així, una categoria acceptable representarà sempre un 33,3 % de l'interval d'ús lingüístic. Com mostren els gràfics 7 i 8, l'experimentació amb nou categories flexibles

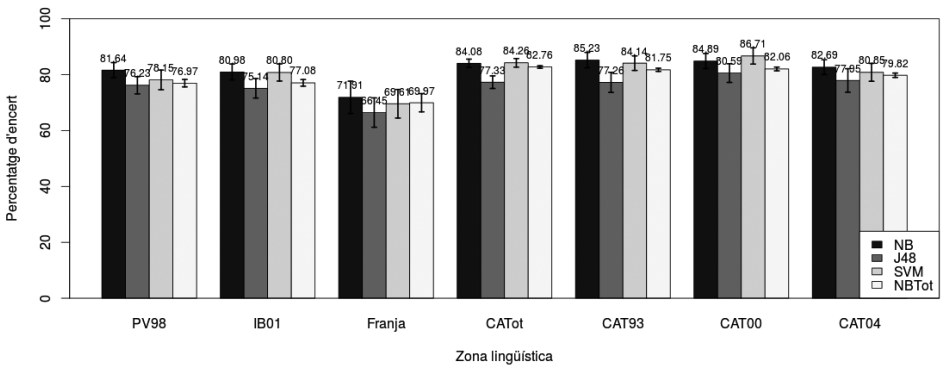
obté percentatges de predicció més alts en la majoria dels casos, amb un guany que arriba als 10 punts percentuals a Catalunya i a la Franja d'Aragó. Cal veure, però, aquesta comparativa més com un exercici matemàtic de validació que no com el millor model possible (encara que els valors del gràfic 8 superen els del gràfic 6), ja que, d'una banda, un model de tres comportaments lingüístics pot ser massa polaritzat en una situació de relació interlingüística entre dues llengües i, d'una altra banda, la definició de fronteres flexibles més amples (33,3 %) garantirà millors resultats però amb una precisió de classificació menor.

GRÀFIC 7
Percentatge d'encert per a tres categories de comportament lingüístic
Predicció del grau d'utilització de la llengua catalana



Els baixos percentatges d'encert mostrats per les barres de NBTot revelen que, al contrari del que s'esperaria en una mostra homogènia, fer servir les dades de tots els territoris i anys com a mostra per a l'entrenament dels models de classificació no millora el resultat de predicció de l'ús lingüístic. Açò és, que el comportament dels individus d'uns territoris no és directament exportable a la resta. La interpretació òbvia és que, tot i que l'ús de la llengua sembla estadísticament ben determinat per les variables proposades, la manera en què s'estableix la relació depèn de l'entorn sociopolític de cada territori, ben diferent en els territoris catalanoparlants. En aquesta línia, Querol (1999 i 2000) ja va anticipar que a Catalunya el nivell alt en la representació de la llengua catalana podria permetre un increment del seu ús, mentre que als altres territoris el fet que aquesta representació siga més baixa fa que la xarxa social passe a ser un factor important.

GRÀFIC 8
Percentatge d'encert per a nou categories flexibles de comportament lingüístic



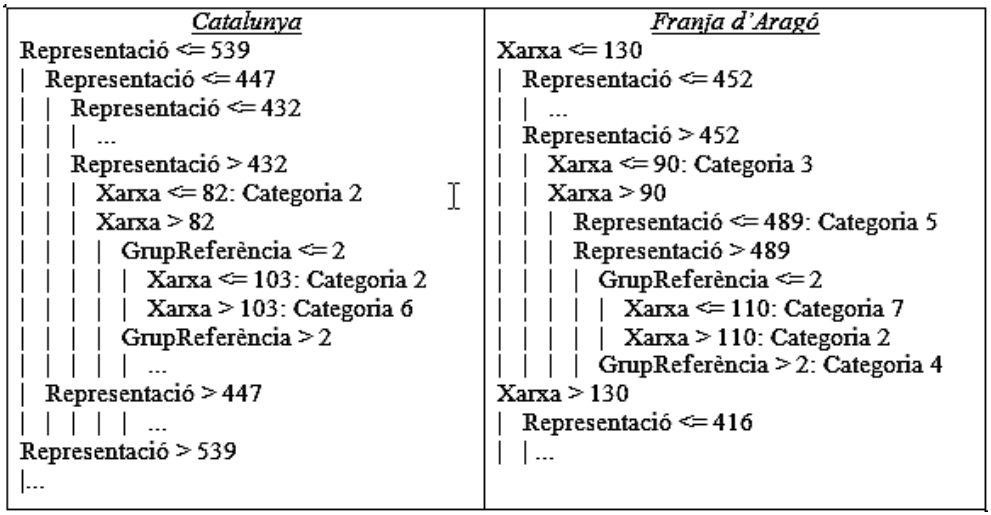
Efectivament, si observem l'estructura interna dels models apresos, com, per exemple, els arbres de decisió J48 del gràfic 9, observem que, en els resultats[9] de Catalunya, la xarxa social apareix en el tercer o quart nivell de la ramificació, és a dir, que la discriminació entre els usos és determinada pels diferents valors de representació de la llengua als nivells més bàsics de selecció. En canvi, en els resultats del País Valencià i de les Illes Balears, la xarxa social apareix ja al segon nivell, després d'una primera discriminació feta en termes de la representació de la llengua. El cas més clar d'aquesta tendència és la Franja d'Aragó, on la xarxa social es troba al primer nivell de l'arbre.

Aspectes similars s'observen en els models bayesians apresos amb NB on les diferents distribucions de les variables dependents (mitjana i desviació estàndard) i els pesos assignats a cada comportament lingüístic defineixen funcions de densitat de probabilitat que reflecteixen aquests mateixos fenòmens per als diferents territoris. I, encara que més difícil de visualitzar, el mateix podem dir dels vectors de suport de les SVM que permeten identificar els individus i comportaments fronterers. Per exemple, de l'anàlisi de les mitjanes obtingudes amb NB podem concloure que als territoris en què la representació del català és típicament alta (525 de mitjana per a Catalunya), aquesta domina l'ús que s'hi fa de la llengua, mentre que allà on la representació és baixa (427 i 439 de mitjana per al País Valencià i les Illes Balears, respectivament), la xarxa social comença a adquirir importància, fins a arribar a convertir-se en el factor més rellevant (423 de mitjana per a la Franja d'Aragó). En definitiva, allà on una llengua deixa de tenir un alt nivell de representació o, dit d'una altra manera, deixa de ser una llengua ben valorada socialment, l'única motivació que resta a la població per a seguir parlant-la és la xarxa social.[10]

9. Els valors numèrics de les variables dependents s'expressen en els intervals definits en la taula 2.

10. En aquest sentit, la reducció de la xarxa social associada a la llengua catalana hi provoca una evident disminució en l'ús (Miralles, 2014) i es retroalimenta amb la caiguda en la representació.

GRÀFIC 9
Exemples d'arbres de decisió per a Catalunya i la Franja



Per acabar, un detall que trobem en els models de classificació generats és que el grup de referència és sovint el criteri que discrimina menys (observeu que apareix en els nodes més interns dels arbres de decisió del gràfic 9). Pensem que els motius poden venir de dues fonts ben diferents: en primer lloc, per la manera com s'han aconseguit les diferents variables i, en segon lloc, per l'edat dels enquestats. Pel que fa a la primera, hem de tenir present que la variable del grup de referència s'aconsegueix només amb la resposta a la pregunta: «A quin grup t'agradaria pertànyer?», a diferència de la representació, que es calcula amb les respostes a 8 escales de 10 preguntes cadascuna, és a dir, a 80 preguntes. Igualment ocorre amb la xarxa social, que s'obté amb les respostes a 4 escales, una de 7 preguntes i les altres 3 de 5 preguntes; per tant, a 22 preguntes. Respecte a la segona, atesa la curta edat dels enquestats (setze anys de mitjana), és molt possible que aquest paràmetre encara no siga rellevant sinó que prendrà un paper més important en edats més tardanes a l'hora d'explicar els canvis en l'ús al llarg dels anys de vida de l'individu que cerca la seua posició en la societat.

5. CONCLUSIONS I TASQUES FUTURES

En aquest article hem comprovat que la classificació supervisada pot servir per a construir models de predicció del grau d'ús del català amb un percentatge d'encert que supera els aconseguits en les investigacions precedents estudiades. A més, els models apresos no s'han de prendre com a caixes negres sinó que ajuden a comprendre el comportament lingüístic dels individus, amb la identificació de quines són les variables més informatives i com aquestes discriminen la pertinença d'un subjecte a una categoria de comportament lingüístic.

El nivell d'encert dels models de classificació canvia depenent de la divisió dels grups d'ús. Per tant, grups més grans donen una visió menys detallada però permeten un alt nivell d'encert. Així, una divisió en deu subgrups amb un marge d'error en l'assignació a un subgrup contigu permet rescabalar el resultat de l'error comès pels pobladors de les fronteres entre els subgrups. Ara bé, el resultat de les enquestes demostra que encasellar els individus en pocs grups no s'ajusta gaire a una realitat complexa en la qual la variable d'ús s'expressa en un espai numèric continu. En definitiva, segurament el sistema requereix un tractament continu més que no pas una subdivisió en grups d'ús. No obstant això, en aquesta primera aproximació al problema mostrem uns resultats inèdits pel que fa a la caracterització dels usos lingüístics que haurien de poder ser exportats a altres casos de sistemes bilingües d'arreu.

Establir la relació universal entre les variables i l'ús independentment del territori i les llengües en contacte hauria de ser l'objectiu últim de treballs com aquest. De moment, els models que es presenten ací proporcionen una foto fixa de persones d'un territori amb una certa edat. Ara bé, tot i que hi ha dades corresponents a anys diferents, és difícil derivar-ne dinàmiques sense tenir enquestes de grups d'edat més avançada, per veure com, a escala individual, les diferents variables proposades com a independents actuen com a motor de canvi en l'ús de cada individu. Per exemple, hom esperaria que el grup de referència d'un individu marque la variació del seu ús amb el temps, ja que aquest individu adoptarà canvis en la seua xarxa social si pretén modificar la seua situació actual. Ara bé, la representació que un individu té respecte a una llengua, la utilitat de la qual pot ser modificada des de les institucions o mitjançant canvis casuals en la xarxa social, podria modificar el seu grup de referència, fet que engegaria el seu moviment pel mapa de l'ús. Aquesta discussió demostra la importància de l'estudi de la dinàmica del sistema, que resta com una tasca futura.

Bibliografia de referència

Boser, Bernhard E.; Guyon, Isabelle M.; Vapnik, Vladimir N. (1992). «A training algorithm for optimal margin classifiers». A: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Nova York: ACM, p. 144-152.

Boudon, Raymond; Cherkaoui, Mohamed; Besnard, Philippe; Lécuyer, Bernard-Pierre (1993). *Dictionnaire de la sociologie*. París: Larousse. [Edició en castellà: *Diccionario de sociología*. Barcelona: Larousse Planeta, 1995]

Calaforra, Guillem (2002). «Quan la sociolingüística abandona l'infantilisme —quatre comentaris sobre una novetat bibliogràfica». *Treballs de Sociolingüística Catalana*, núm. 16, p. 85-92.

Chapelle, Olivier; Haffner, Patrick; Vapnik, Vladimir (1999). «Support vector machines for histogram-based image classification». *IEEE Transactions on Neural Networks*, vol. 10, núm. 5, p. 1055-1064.

Jodelet, Denise (1989). «Représentations sociales: un domaine en expansion». A: Jodelet, Denise (dir.). *Les représentations sociales*. París: Presses Universitaires de France, p. 47-78.

KNERR, Stefan; PERSONNAZ, Leon; DREYFUS, Gerard (1990). «Single-layer learning revisited: a stepwise procedure for building and training a neural network». A: FOGELMAN SOULIÉ, Françoise; HÉRAULT, Jeanny (ed.). *Neurocomputing: Algorithms, architectures and applications*. Berlín: Springer. (NATO ASI Series. Series F, Computer and System Sciences; 68), p. 41-50.

KREMNITZ, Georg (2002). «Recensió de *Cap a un nou marc teòric per a l'estudi de les variables dels processos de substitució lingüística* i *Els valencians i el valencià: Usos i representacions socials* d'Ernest Querol». *Sociolinguistica*, vol. 16, p. 180-182.

MERTON, Robert K; KITT, Alice S. (1950). «Contributions to the theory of reference group behavior». A: MERTON, Robert K.; LAZARSFELD, Paul F. (ed.). *Continuities in social research: Studies in the scope and method of «The American soldier»*. Glencoe, Ill.: Free Press, p. 40-105.

MIRALLES, Clara (2014). «Models dinàmics de competició entre llengües». Treball de fi de grau. València: Universitat de València. Facultat de Física. També disponible en línia a: <http://www.uv.es/perucho/TESI/TGF-CMiralles.pdf> [Consulta: 23 desembre 2015].

MITCHELL, Tom (1997). *Machine learning*. Nova York: McGraw-Hill Higher Education.

QUEROL, Ernest (1999). *Cap a un nou marc teòric per a l'estudi de les variables dels processos de substitució lingüística*. Barcelona: Publicacions Universitat de Barcelona. (Tesis Doctorals Microfitxades; 3568)

— (2000). *Els valencians i el valencià: Usos i representacions socials*. Paiporta: Denes.

— (2002*a*). «A new model to the evaluation of language planning. A case study: Catalonia (1993-2000)». *Sociolinguistica*, vol. 16, p. 129-142.

— (2002*b*). «A new theoretical approach to the study of reversing language shift processes: the catastrophe theory». *7th International Conference on Minority Languages* (Bilbao, 1999). Vitòria: Servicio de Publicaciones del Gobierno Vasco, p. 225-242.

— (2004*a*). «Comparació de resultats empírics sobre representacions socials de les llengües entre les Illes Balears, Catalunya, el País Valencià i Andorra». *Treballs de Sociolingüística Catalana*, núm. 18, p. 43-62.

— (2004*b*). «Empirical corroboration of the catastrophe theory model in Catalonia (1993 and 2000), in the Valencian Country (1998), in the Balearic Islands (2001) and in Andorra (2002)». A: LORENZO SUÁREZ, Anxo M.; RAMALLO, Fernando; RODRÍGUEZ YÁÑEZ, Xoán Paulo (ed.). *Socialización bilingüe e adquisición lingüística bilingüe: Actas do Segundo Simposio Internacional sobre o Bilingüismo* (Vigo, 23-26 octubre 2002). Vigo: Servizo de Publicacións da Universidade de Vigo, p. 1039-1053.

— (2005). «Història sociolingüística recent: Catalunya el 1993, el 2000 i el 2004». *Lengas*, núm. 57: *Brigitte Schlieben-Lange et la sociolinguistique occitane et catalane*, p. 195-218.

REQUENA, Félix (1989). «El concepto de red social». *Revista Española de Investigaciones Sociológicas*, núm. 48, p. 137-152.

VAPNIK, Vladimir (1995). *The nature of statistical learning theory*. Nova York: Springer.

VAPNIK, Vladimir; GOLOWICH, Steven E.; SMOLA, Alex (1997). «Support vector method for function approximation, regression estimation, and signal processing». A: MOZER, Michael C.; JORDAN, Michael I.; PETSCHE, Thomas (ed.). *Advances in neural information processing systems 9*. Cambridge: The Mit Press, p. 281-287.