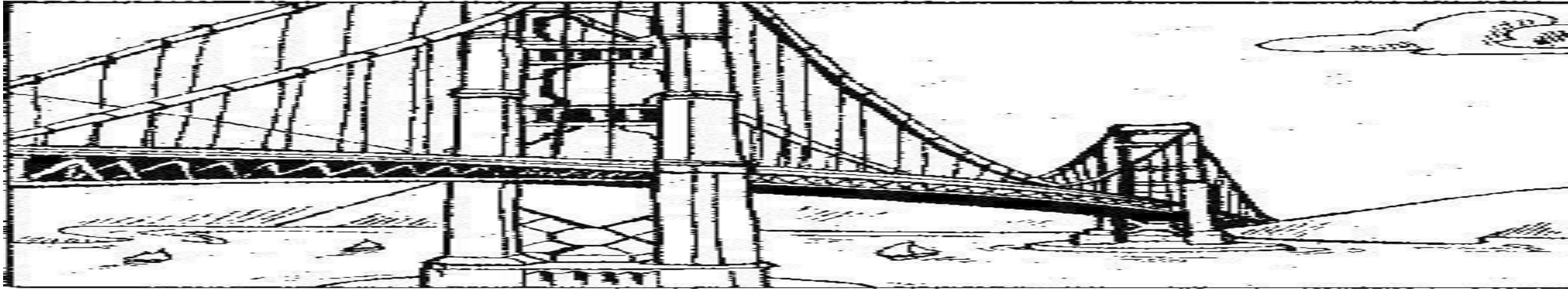


Models dinàmics i models de classificació per a predir l'ús de llengües en competició

Grup de recerca Alcàntera

III Jornades “La recerca sociolingüística en l'àmbit de la llengua catalana”



Grup de recerca Alcàntera

Manel Perucho, Departament d'Astronomia i Astrofísica, Universitat de València

Francisco Grimaldo, Departament d'Informàtica, Universitat de València

Emilia López-Iñesta, Departament d'Informàtica, Universitat de València

Clara Miralles, Màster de Física, Universitat d'Heidelberg

Ernest Querol, Departament d'Arts i Humanitats, Universitat Oberta de Catalunya

TEIXINT PONTS ENTRE SOCIOLINGÜÍSTICA I ELS MÈTODES MATEMÀTICS APLICATS

Grup de recerca Alcàntera: el pont amb la sociolingüística

1. Models dinàmics per a predir l'evolució de l'ús de llengües

1. Models dinàmics per a predir l'evolució de l'ús de les llengües

Miralles, Clara; Perucho, Manel i Querol, Ernest (2015): “Models dinàmics de competició entre llengües aplicats al cas català-castellà al País Valencià”, *Anuari de Psicologia de la Societat Valenciana de Psicologia*, Vol. 16, n. 1, págs. 31-55.

- **Antecedents**

- 2003: Abrams i Strogatz publiquen a [*Nature*](#) sobre **models dinàmics de la desaparició de les llengües**.
- 2005: Mira i Paredes van esmenar el model d'Abrams i Strogatz

- **Mètode**

Recerca inicial de la dinàmica de l'evolució del procés de substitució lingüística al País Valencià a partir de **models matemàtics d'equacions diferencials**.

1. Aizenberg, J., Tkachenko, A., Weiner, S., Addadi, L. & Hendler, G. *Nature* **412**, 819–822 (2001).
2. McPhedran, R. C. *et al. Aust. J. Chem.* **54**, 241–244 (2001).
3. Cattaneo-Vietti, R. *et al. Nature* **383**, 397–398 (1996).
4. Perry, C. C. & Keeling-Tucker, T. J. *Biol. Inorg. Chem.* **5**, 537–550 (2000).
5. Simpson, T. L. *The Cell Biology of Sponges* (Springer, New York, 1984).
6. Sarikaya, M. *et al. J. Mater. Res.* **16**, 1420–1428 (2001).
7. Marcuse, D. *Principles of Optical Fiber Measurements* (Academic, New York, 1981).
8. Clegg, W. J., Kendall, K., Alford, N. M., Button, T. W. & Birchall, J. D. *Nature* **347**, 455–457 (1990).
9. Kamat, S., Su, X., Ballarín, R. & Heuer, A. H. *Nature* **405**, 1036–1040 (2000).
10. Cha, J. N., Stucky, G. D., Morse, D. E. & Deming, T. J. *Nature* **403**, 289–292 (2000).
11. Kroger, N., Lorenz, S., Brunner, E. & Sumper, M. *Science* **298**, 584–586 (2002).

Competing financial interests: declared none.

Linguistics

Modelling the dynamics of language death

Thousands of the world's languages are vanishing at an alarming rate, with 90% of them being expected to disappear with the current generation¹. Here we

should yield the same transition probability as a swap in the fraction of speakers and relative status; thus $P_{xy}(x,s) = P_{yx}(1-x,1-s)$. We also assume that no one will adopt a language that has no speakers ($P_{yx}(0,s) = 0$) or no status ($P_{yx}(x,0) = 0$), and that P_{yx} is smooth and monotonically increasing in both arguments.

These mild assumptions imply that equation (1) generically has three fixed points. Of these, only $x=0$ and $x=1$ are stable. **The model therefore predicts that two languages cannot coexist stably — one will eventually drive the other to extinction.**

To test our model, we collected data on the number of speakers of endangered languages in 42 regions of Peru, Scotland, Wales, Bolivia, Ireland and Alsace-Lorraine, four instances of which are shown in Fig. 1. We fit the model's solutions to the data, assuming transition functions of the forms $P_{yx}(x,s) = cx^a s$ and $P_{xy}(x,s) = c(1-x)^a(1-s)$. Unexpectedly, the exponent a was found to be roughly constant across cultures, with $a = 1.31 \pm 0.25$ (mean $\pm$ standard deviation; further details are available from the authors).

histories of countries where two languages coexist today generally involve split populations that lived without significant interaction, effectively in separate, monolingual societies. Only recently have these communities begun to mix, allowing language competition to begin.

So what can be done to prevent the rapid disintegration of our world's linguistic heritage? The example of Quebec French demonstrates that language decline can be slowed by strategies such as policy-making, education and advertising, in essence increasing an endangered language's status. An extension to equation (1) that incorporates such control on s through active feedback does indeed show stabilization of a bilingual fixed point.

Daniel M. Abrams, Steven H. Strogatz

Department of Theoretical and Applied Mechanics,
Cornell University, Ithaca, New York 14853, USA
e-mail: dma32@cornell.edu

1. Krauss, M. *Language* **68**, 4–10 (1992).
2. Nowak, M. A., Komarova, N. L. & Niyogi, P. *Nature* **417**, 611–617 (2002).
3. Chomsky, N. A. *Syntactic Structures* (Mouton, New York, 1957).
4. Kroch, A. *Lang. Variat. Change* **1**, 199–244 (1989).

90% of them being expected to disappear with the current generation¹. Here we develop a **simple model** of language competition that explains historical data on the decline of Welsh, Scottish Gaelic, Quechua (the most common surviving indigenous language in the Americas) and other endangered languages. A linguistic parameter that quantifies the threat of language extinction can be derived from the model and may be useful in the design and evaluation of language-preservation programmes.

Previous models of language dynamics have focused on the transmission and evolution of syntax, grammar or other structural properties of a language itself²⁻⁷. In contrast, the model we describe here **idealizes languages as fixed, and as competing with each other for speakers**. For **simplicity**, we also assume a **highly connected population, with no spatial or social structure, in which all speakers are monolingual**.

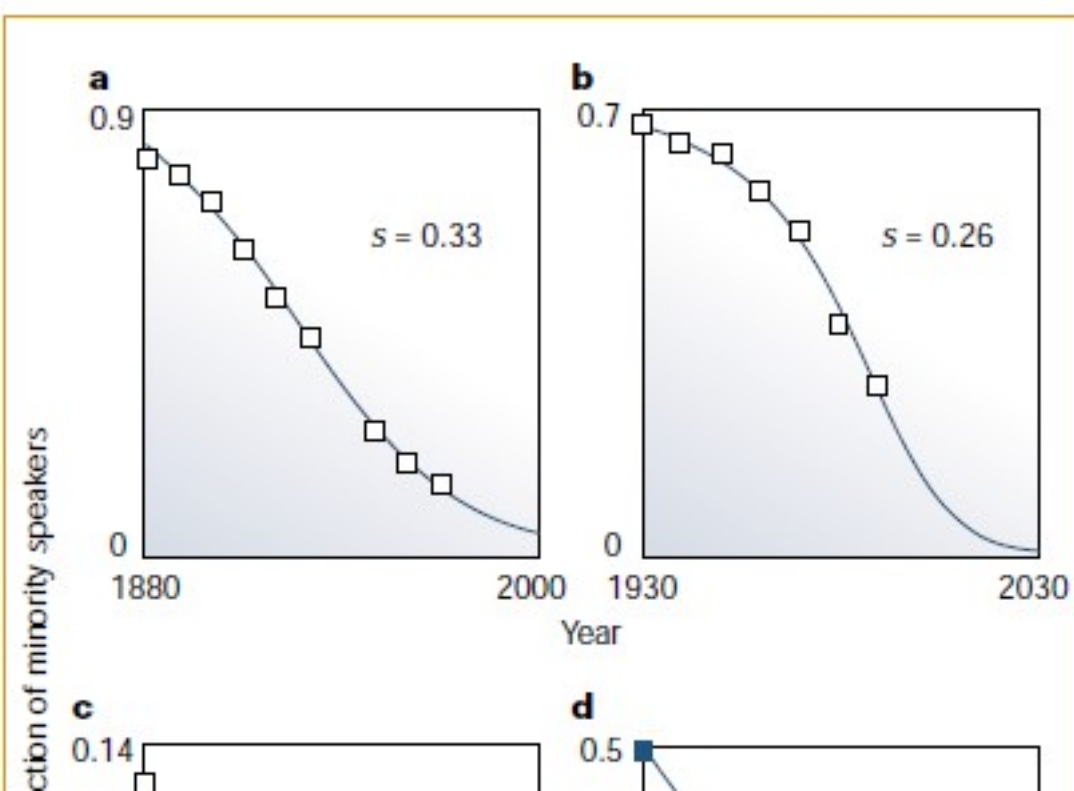
Consider a system of two competing languages, X and Y, in which the attractiveness of a language increases with both its **number of speakers and its perceived status**⁸ (a parameter that reflects the social or economic opportunities afforded to its speakers). Suppose an individual converts from Y to X with a probability, per unit of time, of $P_{yx}(x, s)$, where x is the fraction of the population

further details are available from the authors).

Of the remaining parameters, **status, s , is the most relevant linguistically**; it could serve as a useful measure of the threat to a given language. Quechua, for example, still has many speakers in Huanuco, Peru, but its low status is driving a rapid shift to Spanish, which leads to an unfortunate situation in which a child cannot communicate with his or her grandparents.

Contrary to the model's stark prediction, bilingual societies do, in fact, exist. But the

Figure 1 The dynamics of language death. Symbols show the proportions of speakers over time of: **a**, Scottish Gaelic in Sutherland, Scotland⁹; **b**, Quechua in Huanuco, Peru; **c**, Welsh in Monmouthshire, Wales¹⁰; **d**, Welsh in all of Wales, from historical data¹⁰ (blue) and a single modern census¹¹ (red). Fitted curves show solutions of the model in equation (1), with parameters c , s , a and $x(0)$ estimated by least absolute-values regression. Where possible, data were obtained from several population censuses collected over a long timespan; otherwise, a single



1. New York, 1957).
2. Kroch, A. *Lang. Variat. Change* **1**, 199–244 (1989).
3. Hawkins, J. A. & Gell-Mann, M. *The Evolution of Human Languages* (Addison-Wesley, Reading, Massachusetts, 1992).
4. Niyogi, P. & Berwick, R. C. *Ling. Phil.* **20**, 697–719 (1997).
5. Lightfoot, D. *The Development of Language: Acquisition, Changes and Evolution* (Blackwell, Oxford, 1999).
6. Fishman, J. A. *Reversing Language Shift* (Multilingual Matters, Philadelphia, 1991).
7. Withers, C. W. J. *Gaelic in Scotland 1698–1981: The Geographical History of a Language* (Donald, Edinburgh, 1984).
8. Aitchison, J. & Carter, H. *A Geography of the Welsh Language* (Univ. Wales Press, Cardiff, 1994).
9. UK Office of Population Censuses and Surveys. *1991 Census: Report for Wales, 1* (HMSO, London, 1993).
10. Competing financial interests: declared none.

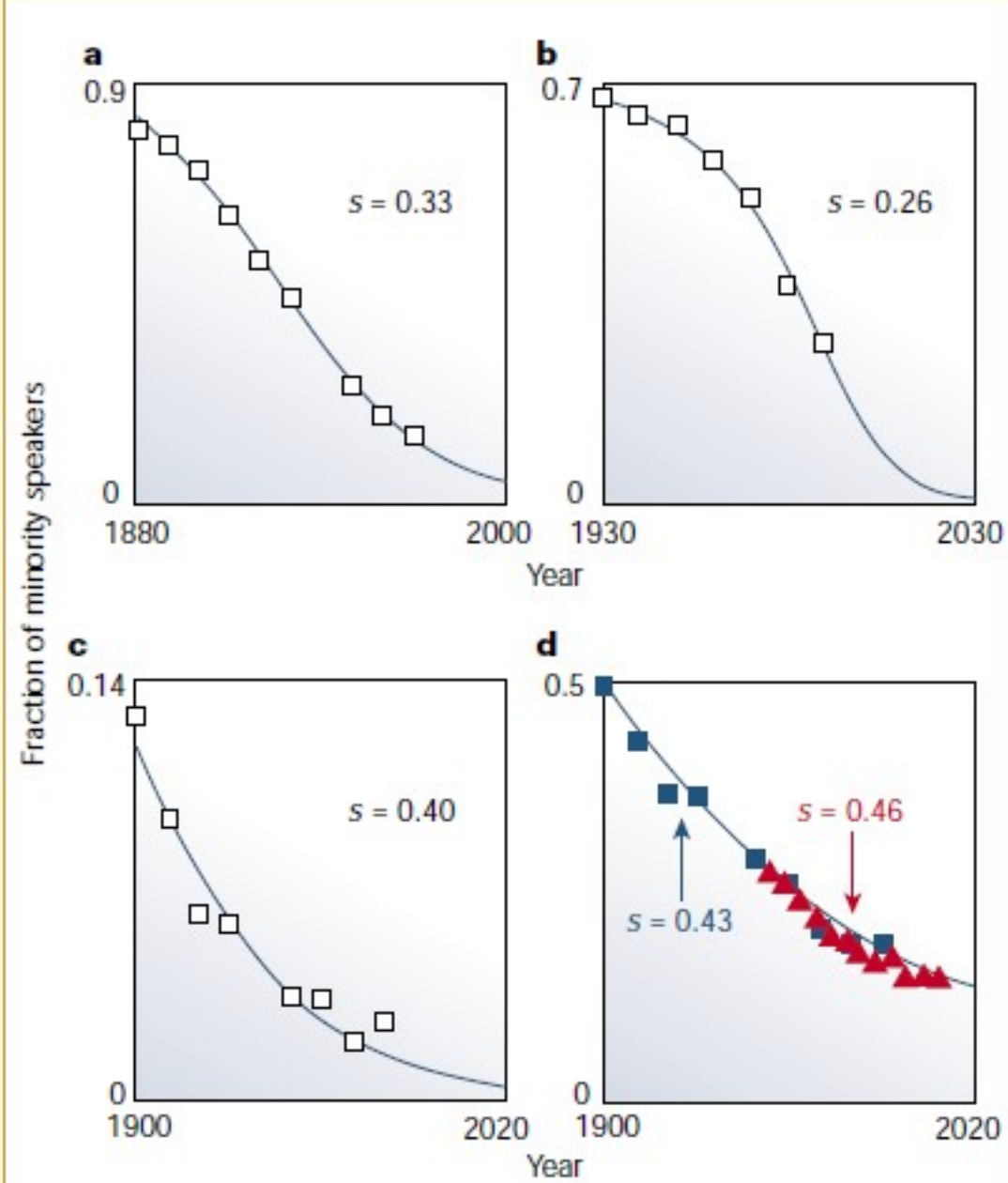
tion of syntax, grammar or other structural properties of a language itself²⁻⁷. In contrast, the model we describe here idealizes languages as fixed, and as competing with each other for speakers. For simplicity, we also assume a highly connected population, with no spatial or social structure, in which all speakers are monolingual.

Consider a system of two competing languages, X and Y , in which the attractiveness of a language increases with both its number of speakers and its perceived status⁸ (a parameter that reflects the social or economic opportunities afforded to its speakers). Suppose an individual converts from Y to X with a probability, per unit of time, of $P_{yx}(x, s)$, where x is the fraction of the population speaking X , and $0 \leq s \leq 1$ is a measure of X 's relative status. A minimal model for language change is therefore

$$\frac{dx}{dt} = yP_{yx}(x, s) - xP_{xy}(x, s) \quad (1)$$

where $y = 1 - x$ is the complementary fraction of the population speaking Y at time t . By symmetry, interchanging languages

Figure 1 The dynamics of language death. Symbols show the proportions of speakers over time of: **a**, Scottish Gaelic in Sutherland, Scotland⁹; **b**, Quechua in Huanuco, Peru; **c**, Welsh in Monmouthshire, Wales¹⁰; **d**, Welsh in all of Wales, from historical data¹⁰ (blue) and a single modern census¹¹ (red). Fitted curves show solutions of the model in equation (1), with parameters c , s , a and $x(0)$ estimated by least absolute-values regression. Where possible, data were obtained from several population censuses collected over a long timespan; otherwise, a single recent census with age-structured data was used (although errors are introduced, the size of which are reflected in the differing fits in **d**). Using the fraction of Catholic masses offered in Quechua in Peru as an indicator, we reconstructed an approximate history of the language's decline.



Les equacions diferencials

Ens permeten:

- **Modelitzar els canvis produïts en una variable** determinada **pels canvis en algun paràmetre** físic rellevant i/o d'altres que hi estiguin lligats.
- El **seguiment continu** dels canvis de les variables en **intervals** de temps **infinitesimals** (concepte de diferencial).

Darreries del s. XX tota una branca de la **física** s'ha llançat a tractar de modelitzar sistemes relacionats amb les **ciències socials**, amb l'ecologia i amb altres àmbits (Strogatz, 1994; Kaplan i Glass, 1995).

Abrams i Strogatz (2003)

- Desenvolupar un model simple de competició entre llengües (tots els parlants són monolingües)
- L'**evolució** de la població de parlants de la llengua A, determinada per **la variable x**, seria:

$$dx/dt = y P_{yx}(x,s) - x P_{xy}(x,s),$$

amb **y el nombre de parlants** de la llengua B, **$P_{yx}(x,s)$ la probabilitat** que els parlants de B passen al grup A, que depén del nombre de parlants d'A (x) i de **l'estatus de la llengua A**, determinat pel **paràmetre s**. Aquest paràmetre reflecteix l'atractiu de la llengua.

- La funció de probabilitat s'expressa de la manera següent: $P_{yx}(x,s) = c x^a s$. En aquesta expressió **c** dóna una idea del **ritme del canvi de llengua entre A i B**, mentre que l'exponent **a** controla la dependència de la probabilitat de canvi amb el nombre de parlants d'A, x.

Model de Paredes i Mira (2005)

Hi inclouen:

- Un tercer grup de parlants: els bilingües.
- Un nou paràmetre: el de semblança, k .

Modula la probabilitat de passar de grup.

- Anàlisi la situació de Galícia: castellà-gallec.

1.1. Models dinàmics: anàlisi de les dades del País Valencià

Dades analitzades

- Serviçi d'Investigació i Estudis Sociolingüístics (SIES) de la Generalitat Valenciana: 1989, 1992, 1995, 2005 i 2010.
- Acadèmia Valenciana de la Llengua (AVL): 2004

Per poder aplicar un model matemàtic d'evolució del nombre de parlants en el temps, ha calgut **transformar els punts (edat, fracció de parlants)** per a un temps fix (any 2004), a punts **(any, fracció de parlants)** per a cada grup d'edat que ens dóna l'enquesta.

Hem considerat que cada grup d'edat representa l'any en el qual els individus tenien de 20 a 30 anys.

1.1. Models dinàmics: anàlisi de les dades del País Valencià

Dades analitzades

- Valors molt semblants entre les enquestes del SIES i l'AVL.
- Paredes i Mira: el paràmetre de semblança entre el castellà i el gallec
 $k = 0.8$

Nosaltres: $K = 0.35$ castellà-valencià

1.1. Models dinàmics: anàlisi de les dades del País Valencià

- **Resultats**

- La tendència decreixent del nombre de parlants és confirmada per les dades del SIES

- **Conclusions principals**

Partint dels intervals de valors possibles dels paràmetres que controlen els models discutim la dependència de l'evolució de l'ús del català amb els paràmetres obtinguts dels ajustos dels models a les dades.

FIGURA 1. *Fraccions de valencianoparlants i castellanoparlants en l'àmbit familiar respecte del temps*

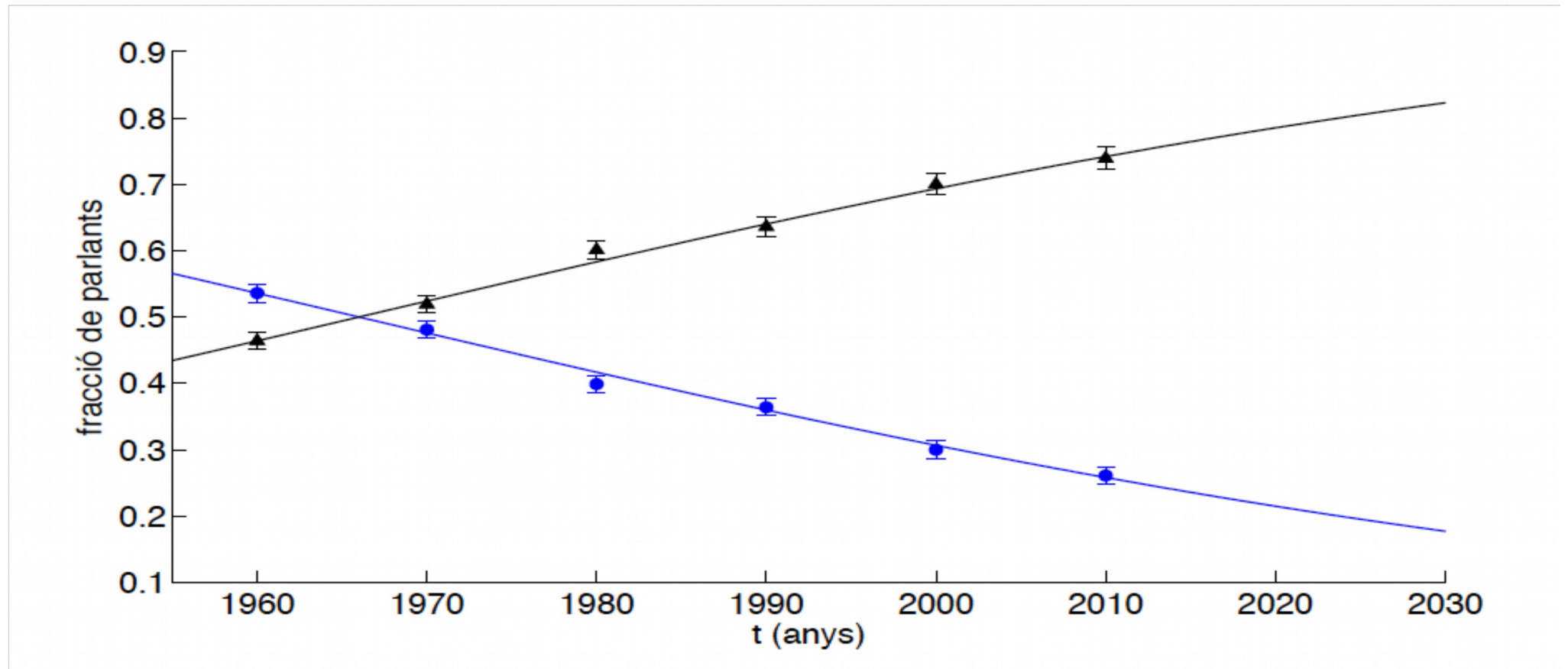
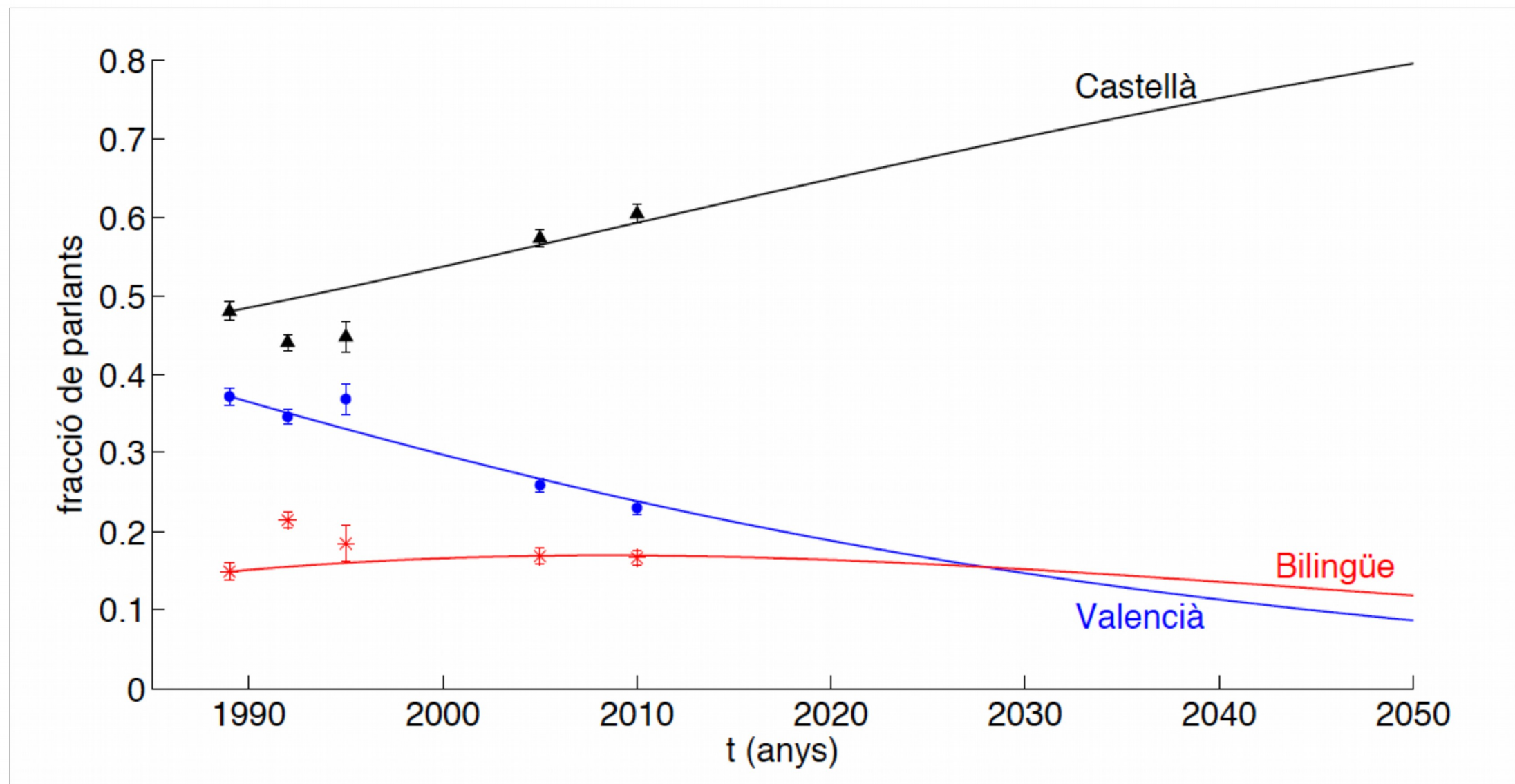


FIGURA 5. *Fracció de parlants en relació amb el temps i classificades en tres grups: valencianoparlants (punts blaus), castellanoparlants (triangles negres) i bilingües (estrelles roges)*

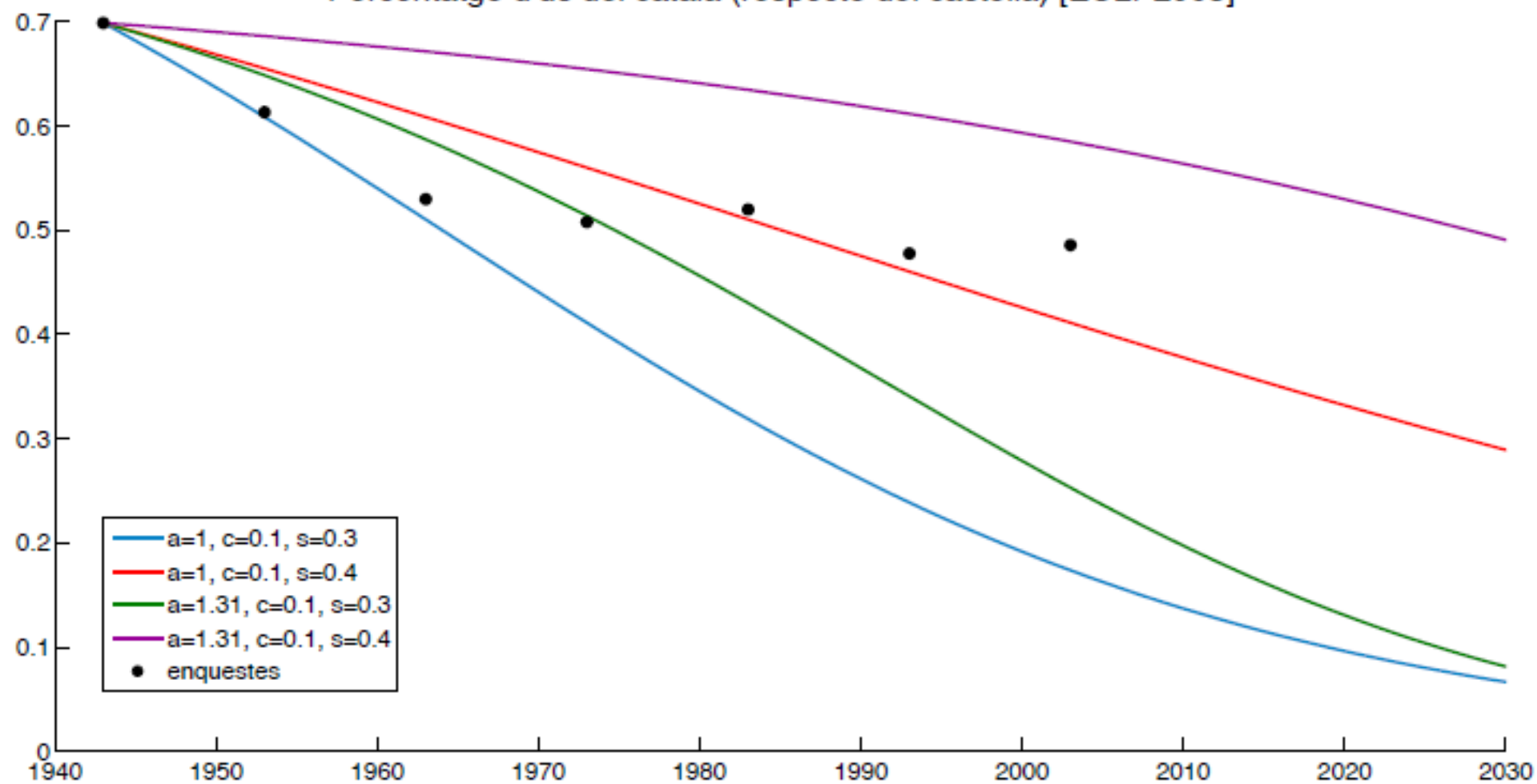


Font: Dades extretes de les enquestes del SIES.

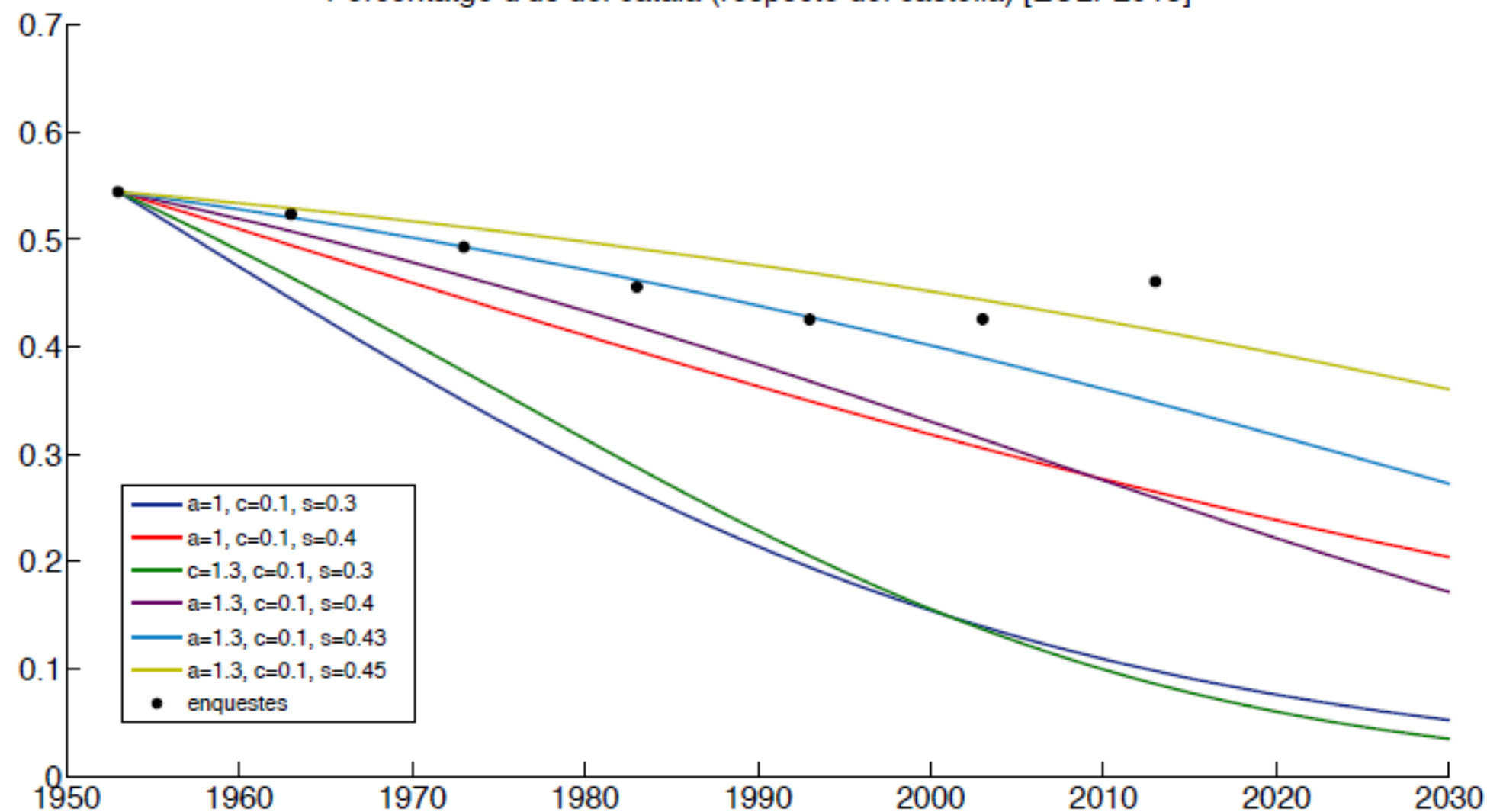
1.3.Podem aplicar el model d'extinció a Catalunya?

El model d'Abrams-Strogatz no es pot **ajustar** a les dades de Catalunya.

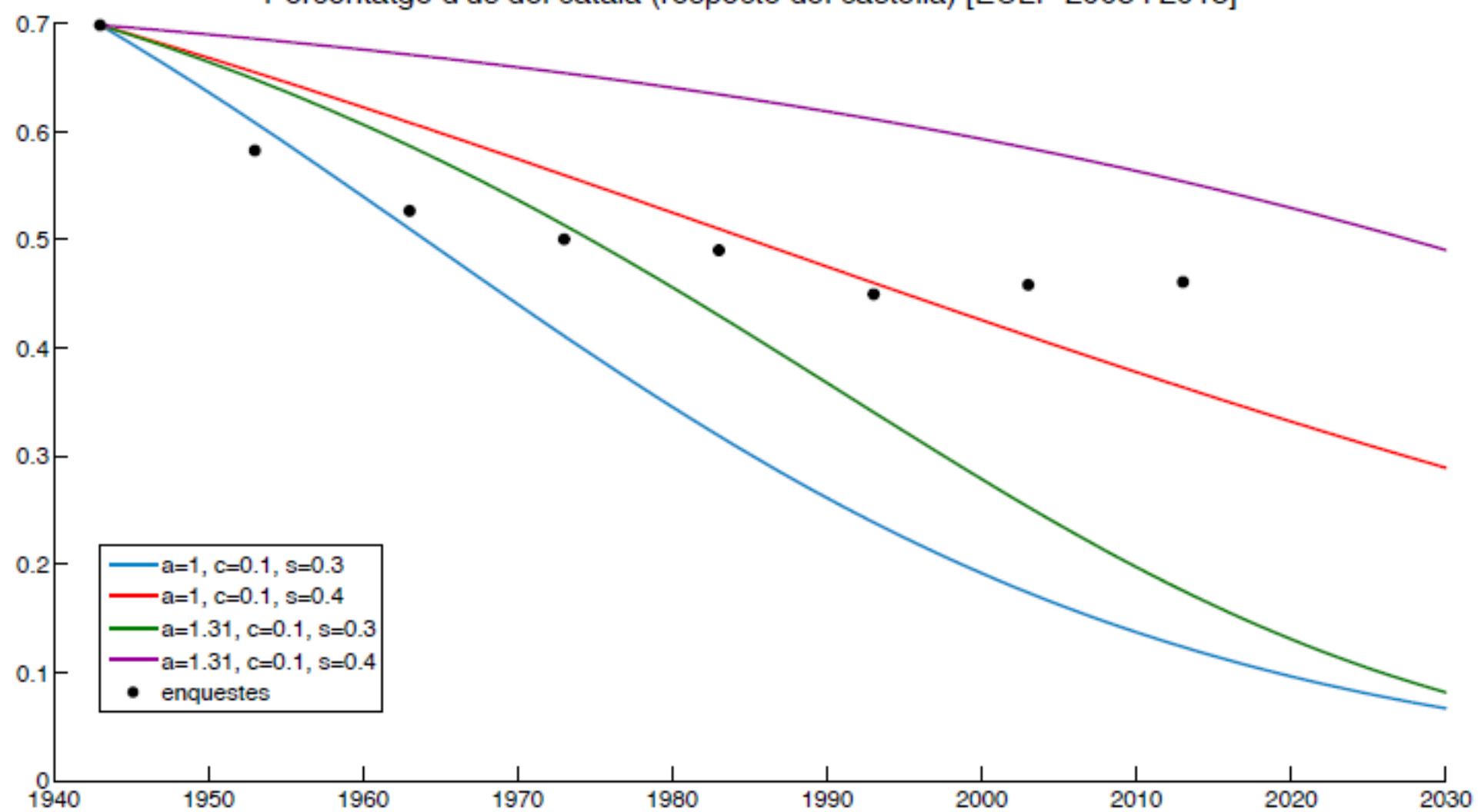
Percentatge d'ús del català (respecte del castellà) [EULP2003]



Percentatge d'ús del català (respecte del castellà) [EULP2013]



Percentatge d'ús del català (respecte del castellà) [EULP 2003 i 2013]



1.3.Podem aplicar el model d'extinció a Catalunya?

Com que és un model d'extinció, **sempre va cap avall** i els paràmetres fan que caigui més dramàticament o menys però no modelen una detenció de la caiguda.

El fracàs de l'ajust mostra que **la dinàmica del català a Catalunya no és la del País Valencià.**

El model és incapaç de reproduir una evolució còncaua (totes les corbes ajustades són convexes).

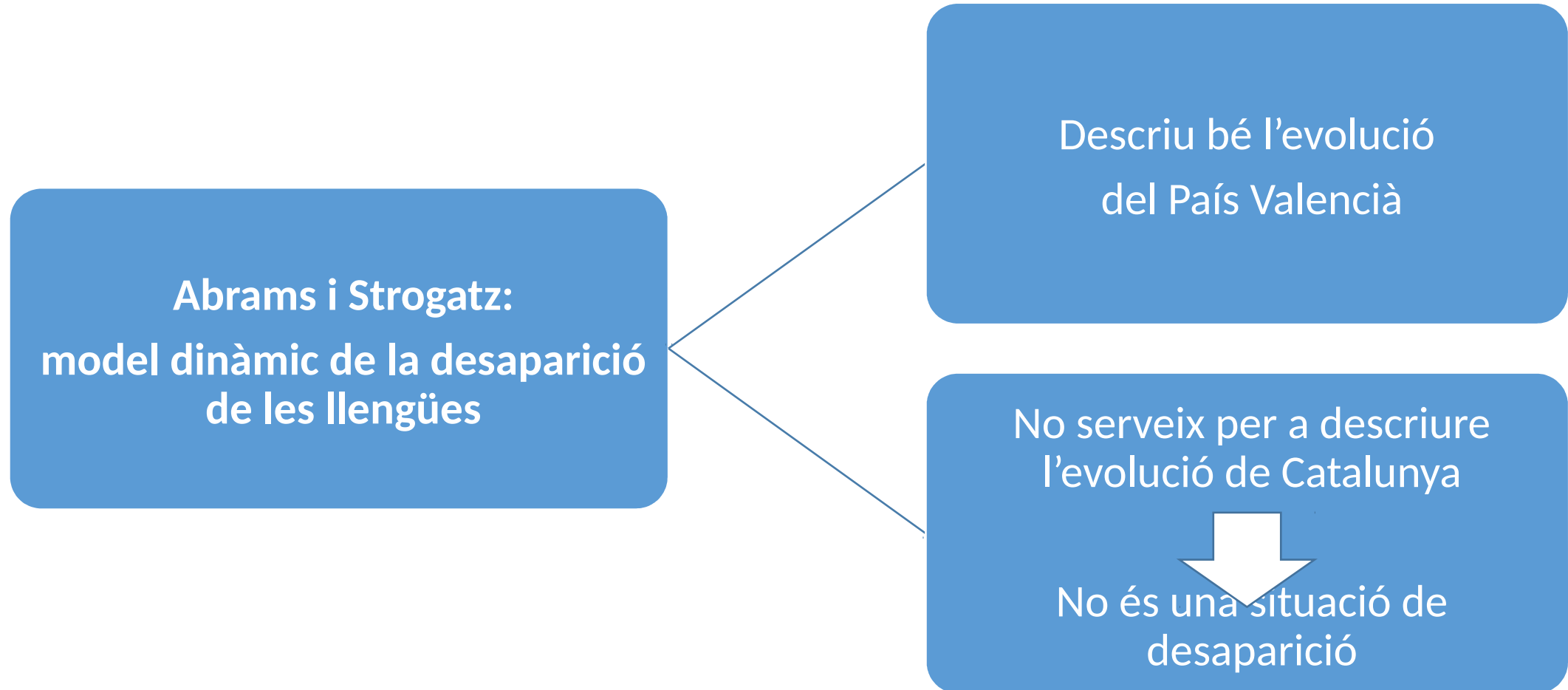
1.3.Podem aplicar el model d'extinció a Catalunya?

El model d'Abrams i Strogatz **sols pot descriure llengües que desapareixen.**

Per tant, mai **no serà capaç de modelitzar tendències de recuperació** com la que s'observa a la gràfica del 2013 (de no ser que s'inverteixin els rols de les llengües en un punt temporal determinat, és a dir, que es passi d'un estatus $s < 0.5$ a un $s > 0.5$).

Això, tanmateix, implicaria un canvi temporal del paràmetre de **l'estatus** (fet desitjable), però que el model no contempla, perquè "**s**" **és un paràmetre fix.**

Conclusions



2. Models de classificació per a predir l'ús de les llengües

2.MODELS DE CLASSIFICACIÓ PER A PREDIR L'ÚS DE LES LLENGÜES

Grimaldo, f., López-Iñesta, E, Perucho, M. i Querol, E. (2016): "Predicció de l'ús del català mitjançant la classificació supervisada", *Treballs de Sociolingüística Catalana*, 26. En premsa.

Model de variables que influeixen els usos lingüístics

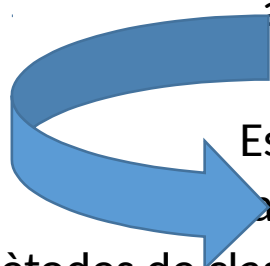


Podem millorar aquest model?



APRENTATGE AUTOMÀTIC (*Machine Learning*) = mètodes d'una àrea de la intel·ligència artificial

Implementació de **mètodes computacionals** que permeten **induir models de coneixement** a partir d'informació que prové de dades d'exemple disponibles

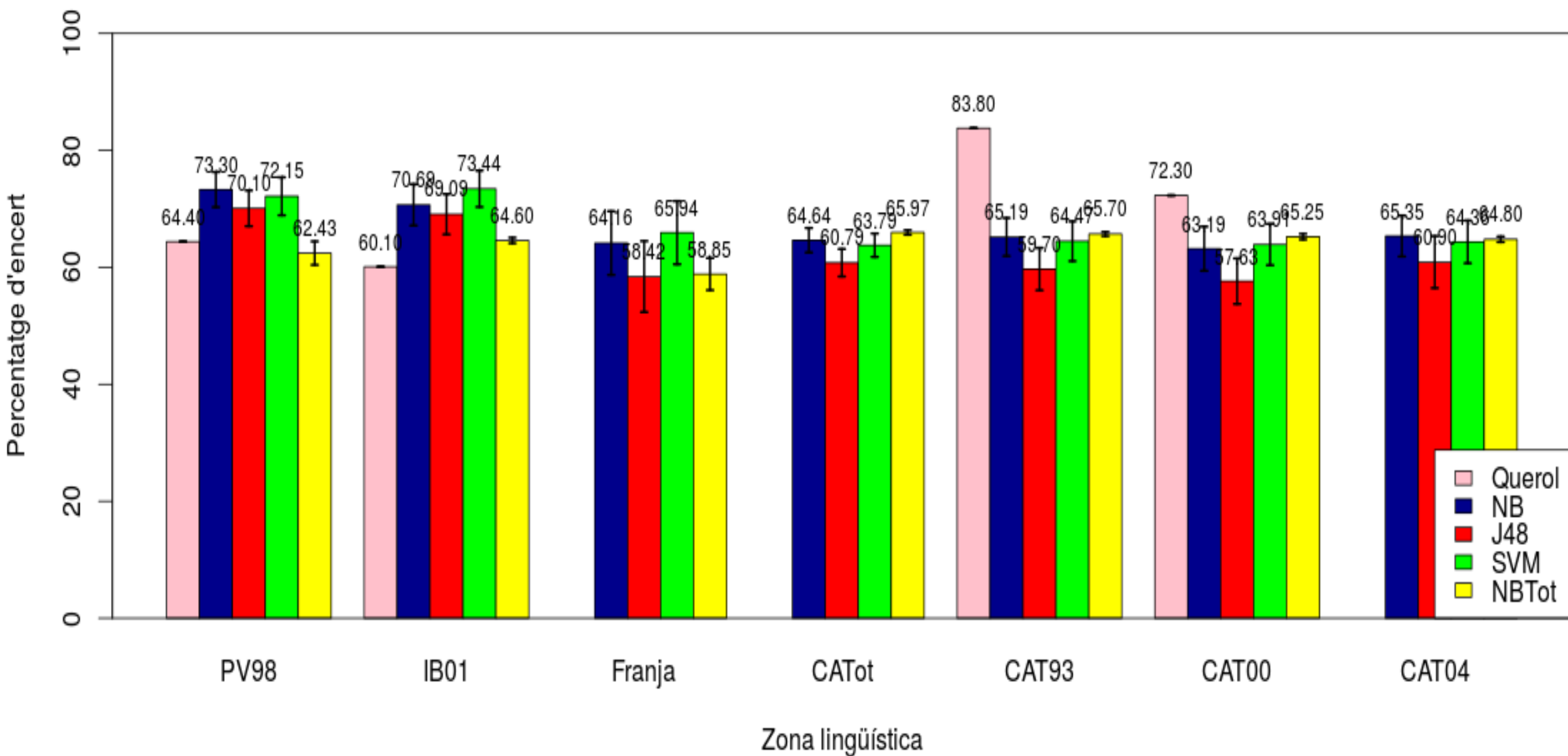


Escatir si algun d'ells millora la predicció del grau d'utilització de la llengua catalana aconseguida fins ara

Hi hem fet servir tres mètodes de classificació supervisada:

- Naive Bayes
- Arbres de decisió
- Màquines de vector suport

Predicció grau d'utilització de la llengua catalana



Què ens han aportat els models de classificació supervisada?

- Comprovar el nivell de predicció del nostre **model teòric** i la seua **validesa en diferents contextos sociolingüístics**.
- Serveix 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.
- Estableix quines són les **variables més informatives**.
- Ens ajuda a resoldre el problema metodològic de **la divisió en grups lingüístics** i palesa que l'ús és un **sistema continu**.

3. *Altres recerques*

3. Altres recerques

Ja acabades:

- Estudi dels models de classificació a la Catalunya del Nord.

En curs:

- Aplicació dels models de classificació a Andorra i a l'Alguer.
- Aplicació dels models d'Abrams i Strogatz i de Paredes i Mira a dades del País Basc
- Modelització de les dades de les Enquestes d'Usos Lingüístics de la Població (2003, 2008 i 2013) que expliquin l'evolució dels usos a Catalunya.

Futures

- Modelització de les dades de les Enquestes d'Usos Lingüístics de la Població de la resta de territoris (2003 i 2004 i les recents d'ara).

Gràcies per l'atenció!



Molt bon profit!