XXVIII Congresso Internazionale di Lingüística e Filologia Romanza
Roma 18-23 luglio 2016

Noves estratègies per a l'anàlisi de la variació lingüística a partir d'un corpus oral dialectal

Esteve Clua
U Pompeu Fabra
FFI2013-46987-C3-3-P

Maria-Rosa Lloret
U Barcelona
FFI2013-46987-C3-1-P

Miquel Salicrú
U Barcelona
FFI2013-46987-C3-3-P

upf. Universitat Pompeu Fabra Barcelona
IULA Institut Universitari de Lingüística Aplicada
Universitat de Barcelona

Esquema de la presentació

1. El COD2 (Corpus Oral Dialectal)
2. Anàlisi lingüística
3. Variació espacial i temporal
4. Anàlisi dialectomètrica
5. Conclusions

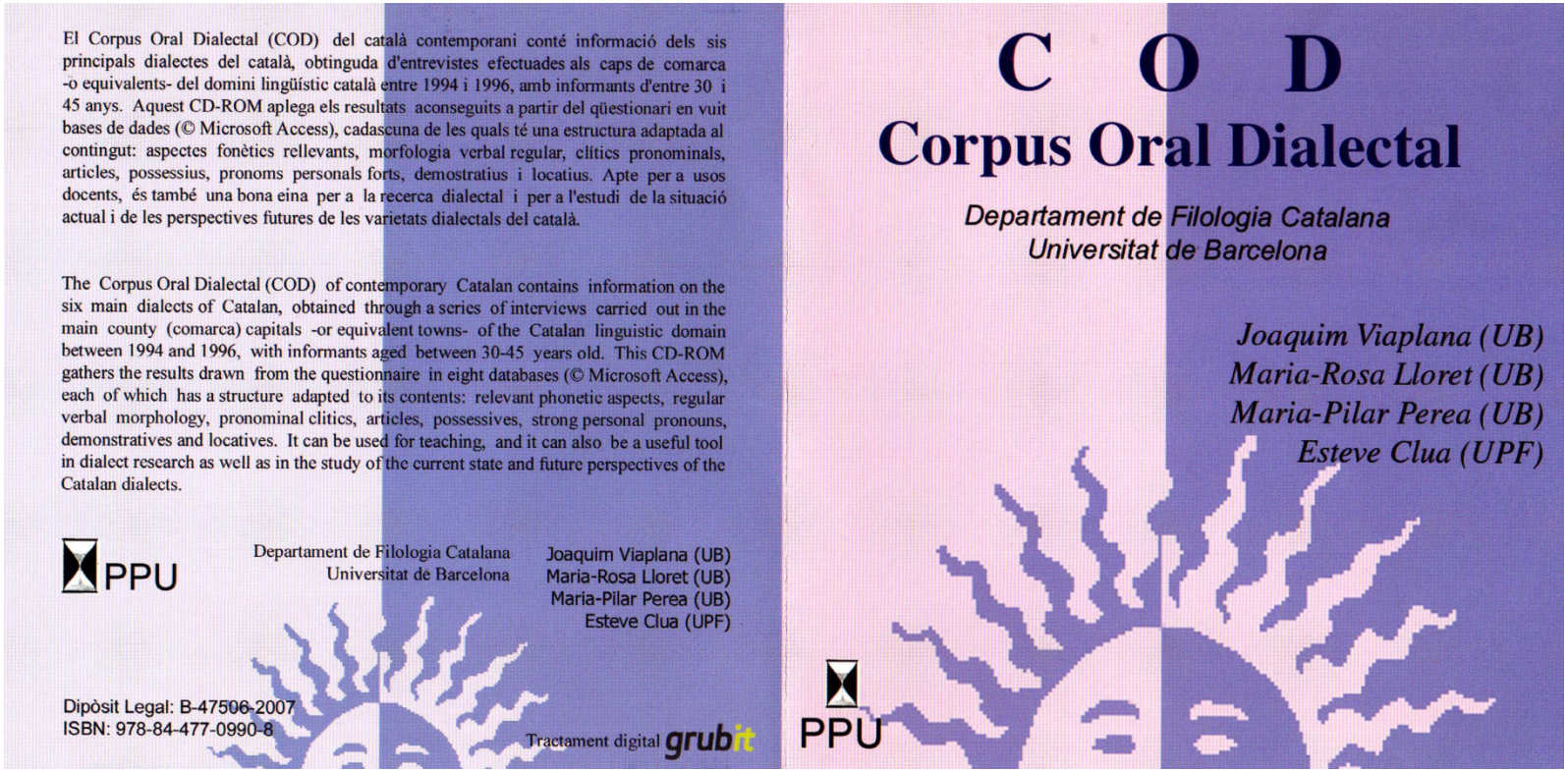
1. El COD2 (**C**orpus **O**ral **D**ialectal del català contemporani, fase 2)

Corpus Oral Dialectal 2 del català contemporani (COD2)

- El COD2 és un corpus de les característiques fòniques i morfològico-flexives de les varietats geogràfiques del català, actualitzat el 2014 a partir d'un corpus de 1994 (COD).
- La finalitat del COD2 és contribuir a ampliar el coneixement sobre la **variació lingüística** en general i, en particular, sobre la **distància** entre varietats lingüístiques, des d'una doble perspectiva: l'**espacial** i la **temporal** (**canvi lingüístic** entre 1994 i 2014).

Resultats anteriors (COD)

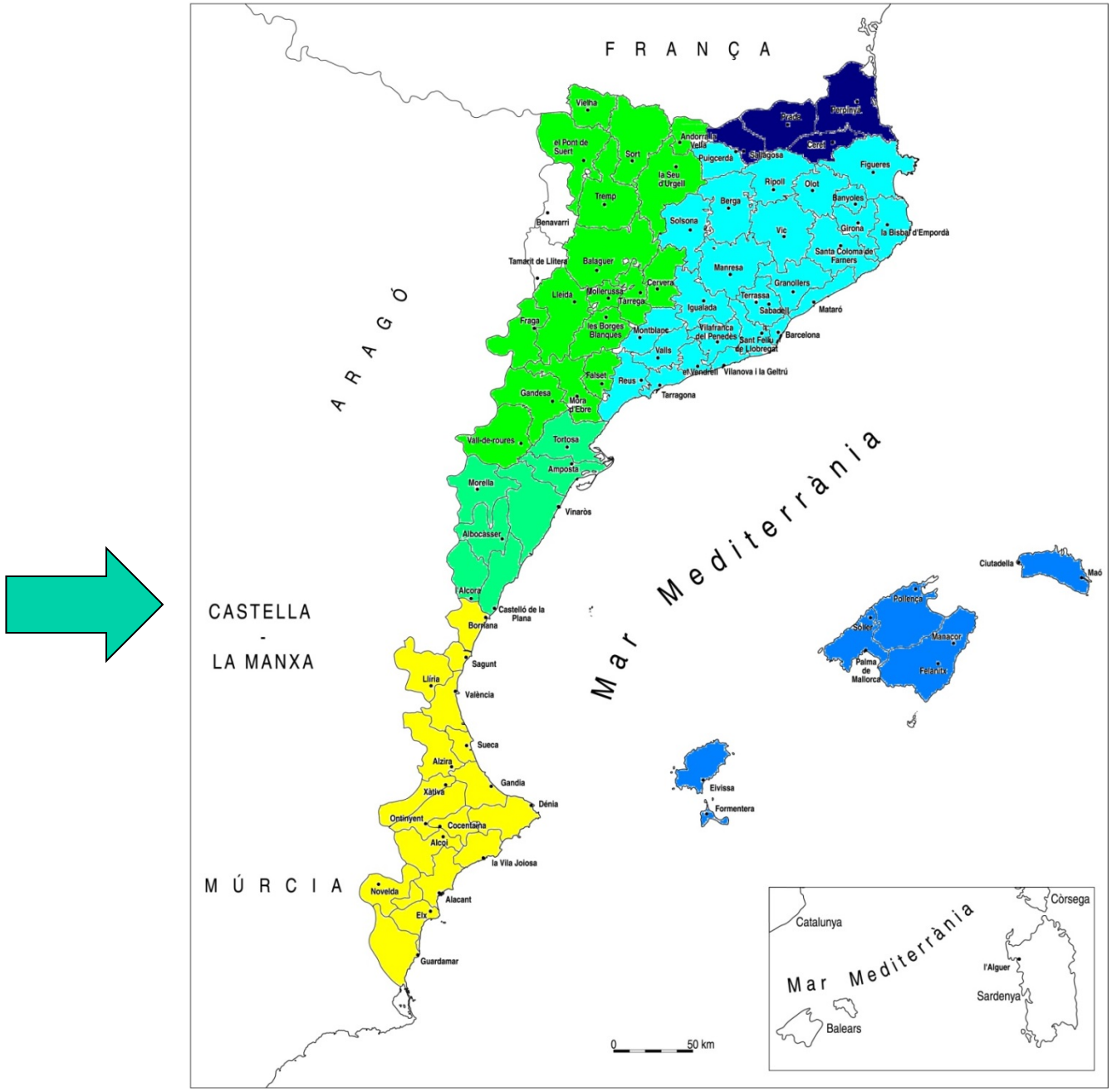
- http://www.ub.edu/lincat



COD2. Característiques

- Qüestionari de 700 ítems (600 en el COD)
- Textos espontanis (10 minuts)
- 82 capitals de comarca de l'àmbit lingüístic del català
- 3 informants com a mínim, de 30 a 45 anys i de nivell cultural mitjà

82 capitals de comarca (o equivalents)



Àmbits del qüestionari

○ Articles:	34 preguntes	
○ Clítics pronominals:	78 preguntes	(72)
○ Demostratius:	17 preguntes	
○ Possessius:	22 preguntes	
○ Locatius:	3 preguntes	
○ Pronoms personals:	8 preguntes	
○ Flexió verbal:	250 preguntes	
○ Flexió nominal:	20 preguntes	(4)
○ Lèxic i fonètica:	241 preguntes	(180)

Text lliure 10 minuts

En aquesta presentació, destacarem les novetats metodològiques del COD2 en l'anàlisi de la variació dialectal quant a:

- L'anàlisi lingüística (v. § 2)
- La variació espacial i temporal (v. § 3)
- L'anàlisi dialectomètrica (v. § 4)

2. Anàlisi lingüística

Mètode COD: Anàlisi generativa clàssica (regles)

- **Diferències subjacents** (computen individualment):

  Proclític de 3a pers. masc. sing. acusatiu:

  el /l+Ø/ - lo /l+o/: 1 diferència (morf de 'masculí': Ø - o)

- **Diferències superficials** (computen en forma de **regles**, del número de canvis):

  Proclític de 1a pers. sing.: /m/: [əm], [em], [am]; [mə], [me], [ma]

[əm] - [em]: 1 diferència [əm] - [am]: 1 diferència [em] - [am]: 1 diferencia	[mə] - [me]: 1 diferència [mə] - [ma]: 1 diferència [me] - [ma]: 1 diferència
[əm] - [mə]: 1 diferència [em] - [me]: 1 diferència [am] - [ma]: 1 diferència	[əm] - [me]: 2 diferències [əm] - [ma]: 2 diferències [em] - [ma]: 2 diferències, etc.

Mètode COD2: Teoria de l'optimitat (restriccions)

- **Diferències subjacents** (computen individualment):

  Proclític de 3a pers. masc. sing. acusatiu:

  el /l+Ø/ - lo /l+o/: 1 diferència (morf de 'masculí': Ø - o)

- **Diferències superficials** (computen en forma de **distàncies** entre les **restriccions** responsables de les diferències):

  Proclític de 1a pers. sing.:

  /m/: [əm], [em], [am] ; [mə], [me], [ma]

Anàlisi en teoria de l'optimitat

- Restriccions que afavoreixen les vocals menys sonants (menys prominents) en posició àtona; ordenació universal:

  *ÀTON/a >> *ÀTON/e >> *ÀTON/ə

- Restriccions que afavoreixen que els nuclis (N) sil·làbics, en general, siguin més sonants; ordenació universal:

  *N/ə >> *N/e >> *N/a

[ə]	[a]	[e]
1. ***ÀTON/a**	1. ***N/ə**	1. ***ÀTON/a, *N/ə**
2. ***ÀTON/e**	2. ***N/e**	2. ***ÀTON/e, *N/e**
3. ***ÀTON/ə, *N/ə**	3. ***N/a, *ÀTON/a**	3. *ÀTON/ə, *N/a
4. *N/e	4. *ÀTON/e	
5. *N/a	5. *ÀTON/ə	

[əm] – [am]: 12 (abans 1 sola diferència)

[əm] – [em]: 6 (abans 1 sola diferència)

[em] – [am]: 6 (abans 1 sola diferència)

	1	2	3	4	5
	*ÀTON/a	*ÀTON/e	*ÀTON/ə, *N/ə	*N/e	*N/a
☞ [ə]			* *		
[a]	*!				*
[e]		*!		*	

2+2+2+2+2+2 = 12

	1	2	3	4	5
	*N/ə	*N/e	*N/a, *ÀTON/a	*ÀTON/e	*ÀTON/ə
[ə]	*!				
☞ [a]			* *		*
[e]		*!		*	

	1	2	3	4	5
	*ÀTON/a	*ÀTON/e	*ÀTON/ə, *N/ə	*N/e	*N/a
☞ [ə]			* *		
[a]	*!				*
[e]		*!		*	

2+2+2 = 6

	1	2	3
	*ÀTON/a, *N/ə	*ÀTON/e, *N/e	*ÀTON/ə, *N/a
[ə]	*!		*
[a]	*!		*
☞ [e]		* *	

	1	2	3	4	5
	*N/ə	*N/e	*N/a, *ÀTON/a	*ÀTON/e	*ÀTON/ə
[ə]			* *		
☞ [a]	*!				*
[e]		*!		*	

2+2+2 = 6

	1	2	3
	*ÀTON/a, *N/ə	*ÀTON/e, *N/e	*ÀTON/ə, *N/a
[ə]	*!		*
[a]	*!		*
☞ [e]		* *	

[əm] – [am]: 12 (abans 1 sola diferència)

[əm] – [em]: 6 (abans 1 sola diferència)

[em] – [am]: 6 (abans 1 sola diferència)

Investigar quines conseqüències pot tenir el tipus d'anàlisi lingüística a l'hora de computar les distàncies lingüístiques.

3. Variació espacial i temporal

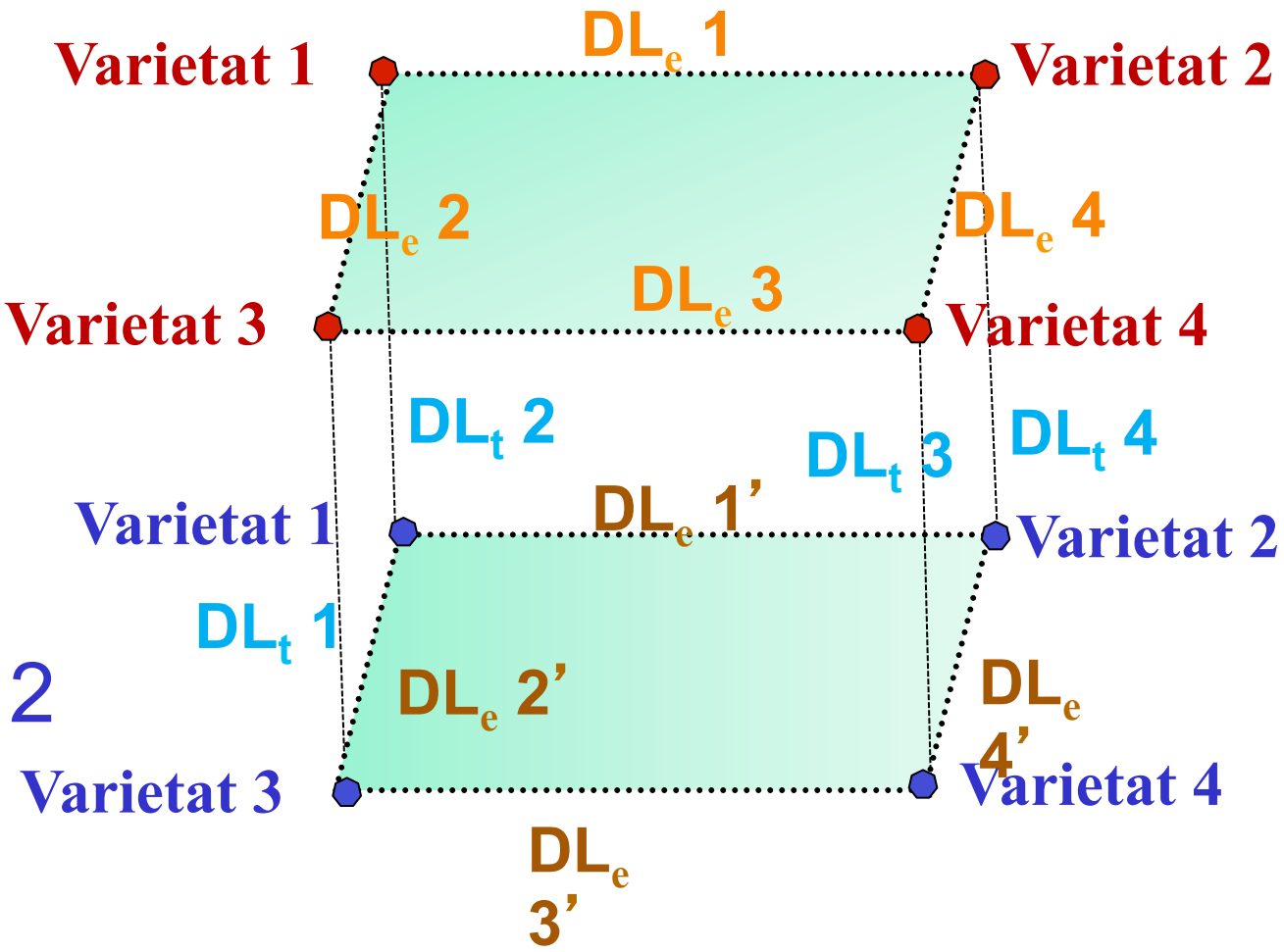
Concepte de Distància Lingüística (DL)

- Quantificació de les similituds o de les diferències lingüístiques entre individus, poblacions o grups de poblacions

Distàncies Lingüístiques: DL_e ESPACIAL
DL_t TEMPORAL

1994 COD

2014 COD 2



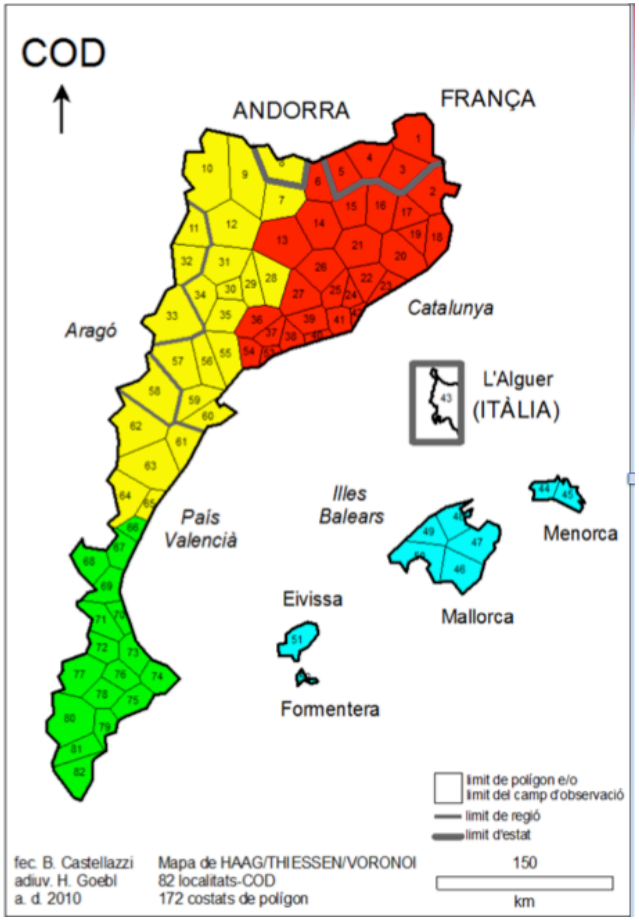
4. ANÀLISI DIALECTOMÈTRICA

Mètodes deterministes vs. mètodes fuzzy-probabilístics

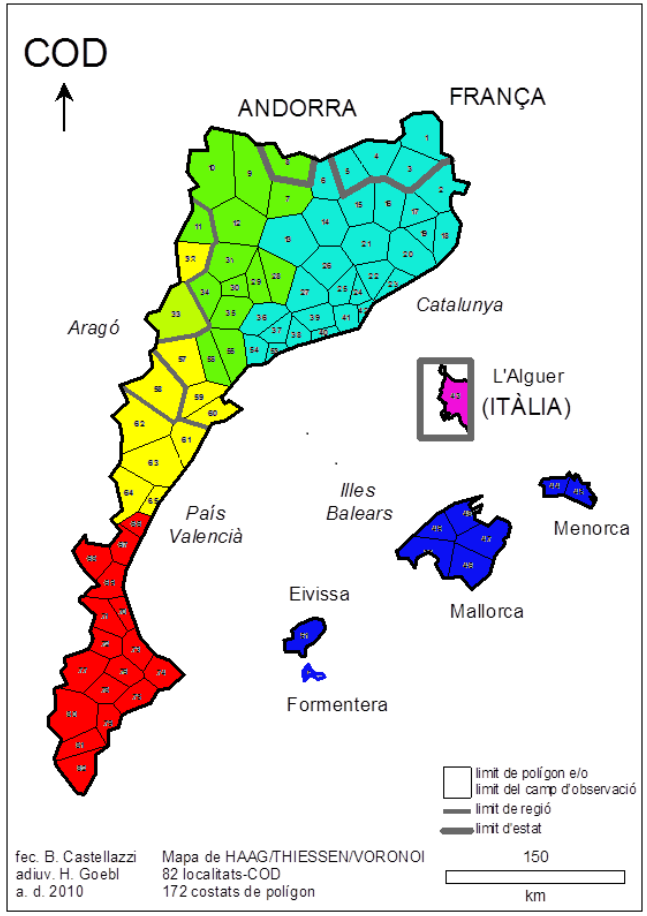
- Els mètodes deterministes s'orienten cap a la classificació de les poblacions (o varietats) en grups dialectals, mentre que amb mètodes fuzzy o probabilístics la classificació de les poblacions es complementa amb la versemblança de pertinença a cada grup.

- Així, mentre que la classificació determinista ens condueix a la representació de grups disjunts, la representació fuzzy ens proporciona una estructura de grups més sòlida i posa de manifest les poblacions frontera.

Classificació Determinista



Classificació FUZZY C-means



Classificació Fuzzy

	g-1	g-2	g-3	g-4	g-5	g-6
Amposta	0	0	0	0,97	0,03	0
Andorra la Vella	0	0	0	0	1	0
Balaguer	0	0	0	0	1	0
Benavarri	0,01	0	0,03	0,1	0,86	0
Cervera	0	0	0	0	1	0
El Pont de Suert	0	0	0	0,31	0,68	0
Falset	0	0	0	0	1	0
Fraga	0	0	0	0,36	0,64	0
Gandesa	0	0	0	0,99	0,01	0
La Seu d'Urgell	0	0	0	0	1	0
Les Borges Blanques	0	0	0	0	1	0
Lleida	0	0	0	0	1	0
Mollerussa	0	0	0	0	1	0
Móra d'Ebre	0	0	0	0	1	0
Sort	0	0	0	0,1	0,9	0
Tamarit de Llitera	0	0	0	0,95	0,05	0
Tàrrega	0	0	0	0	1	0
Tortosa	0	0	0	0,99	0,01	0
Tremp	0	0	0	0,01	0,99	0
Vall-de-roures	0	0	0	1	0	0
Albocàsser	0	0	0	1	0	0
Castelló de la Plana	0	0	0,01	0,99	0	0
L'Alcora	0	0	0	1	0	0
Morella	0	0	0	0,99	0	0
Vinaròs	0	0	0,06	0,91	0,02	0

Classificació Fuzzy

Exemple:

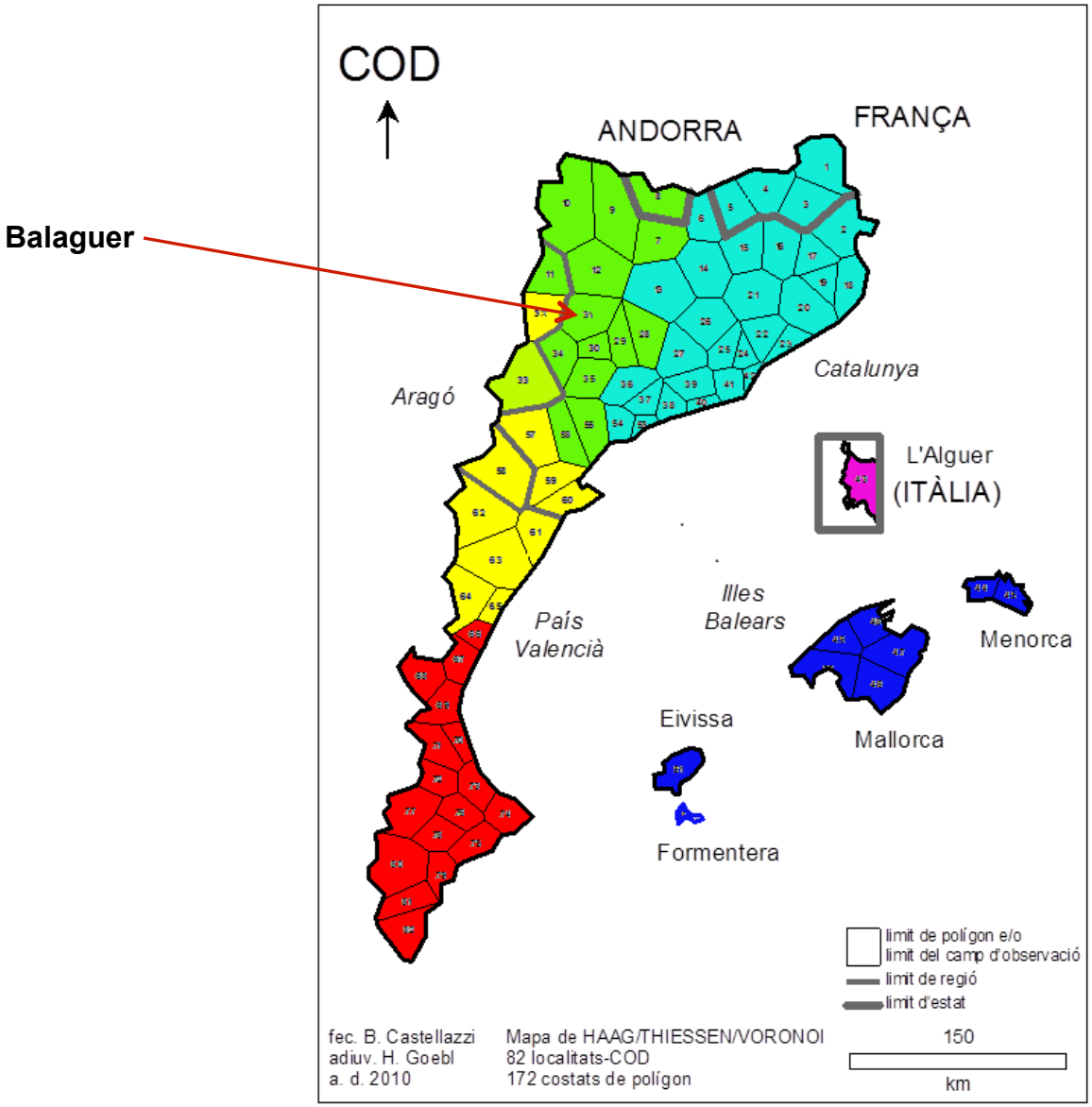
- **Balaguer (31)**

  Probabilitat de pertinença al G5 (Nord-occidental): 1

  Probabilitat de pertinença a altres grups: 0

  Població que pertany nítidament al G5

Classificació FUZZY



Classificació Fuzzy

	g-1	g-2	g-3	g-4	g-5	g-6
Amposta	0	0	0	0,97	0,03	0
Andorra la Vella	0	0	0	0	1	0
Balaguer	0	0	0	0	1	0
Benavarri	0,01	0	0,03	0,1	0,86	0
Cervera	0	0	0	0	1	0
El Pont de Suert	0	0	0	0,31	0,68	0
Falset	0	0	0	0	1	0
Fraga	0	0	0	0,36	0,64	0
Gandesa	0	0	0	0,99	0,01	0
La Seu d'Urgell	0	0	0	0	1	0
Les Borges Blanques	0	0	0	0	1	0
Lleida	0	0	0	0	1	0
Mollerussa	0	0	0	0	1	0
Móra d'Ebre	0	0	0	0	1	0
Sort	0	0	0	0,1	0,9	0
Tamarit de Llitera	0	0	0	0,95	0,05	0
Tàrrega	0	0	0	0	1	0
Tortosa	0	0	0	0,99	0,01	0
Tremp	0	0	0	0,01	0,99	0
Vall-de-roures	0	0	0	1	0	0
Albocàsser	0	0	0	1	0	0
Castelló de la Plana	0	0	0,01	0,99	0	0
L'Alcora	0	0	0	1	0	0
Morella	0	0	0	0,99	0	0
Vinaròs	0	0	0,06	0,91	0,02	0

Classificació Fuzzy

Exemple:

- **Balaguer (31)**

  Probabilitat de pertinença al G5 (Nord-occidental): 1

  Probabilitat de pertinença a altres grups: 0

  Població que pertany nítidament al G5

- **Fraga (33)**

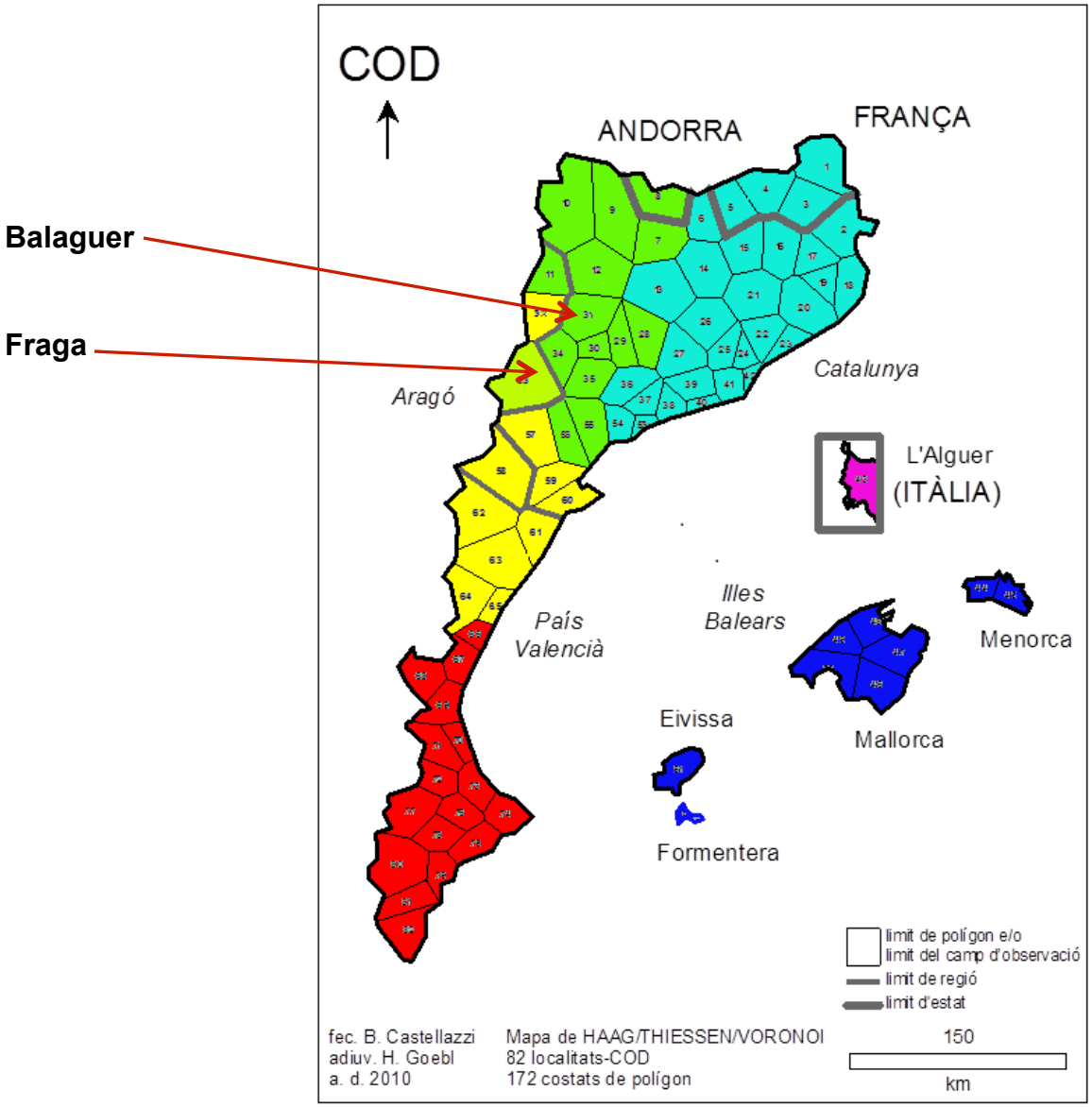
  Probabilitat de pertinença al G5 (Nord-occidental): 0,64

  Probabilitat de pertinença al G4 (Tortosí/Franja): 0,36

  Probabilitat de pertinença a altres grups: 0

  Població frontera entre els grups G5/G4

Classificació FUZZY



Classificació i caracterització de varietats dialectals

- Un dels objectius de la nostra anàlisi dialectomètrica se centra a identificar les poblacions i les formes que permeten la caracterització de les varietats lingüístiques resultants d'un procés de classificació determinista.
- Quant a les poblacions, l'interès s'orienta a identificar-ne aquelles que constitueixen la referència de cada grup (poblacions centrals o de referència), i
- quant a les formes, ens interessa identificar les més informatives, tant a nivell global com per cadascun dels grups; en aquest últim cas es tractarà de determinar les formes més representatives i les més distintives.

Poblacions centrals o de referència (1)

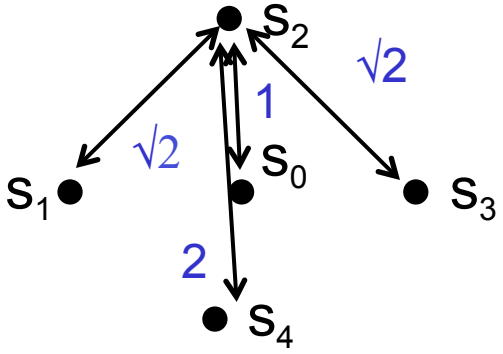
Quan una població es troba en un extrem del grup (d'acord amb la DL) la seva distància en relació amb les altres poblacions és alta.

Per contra, quan la població és central, la distància que la separa de les altres poblacions es redueix.

En conseqüència la població central o de referència és la que presenta la mínima distància mitjana en relació amb les altres poblacions.

Poblacions centrals o de referència (2)
Quina és la S_{ref}?

Si tenim un grup de poblacions: s_0, s_1, s_2, s_3, s_4

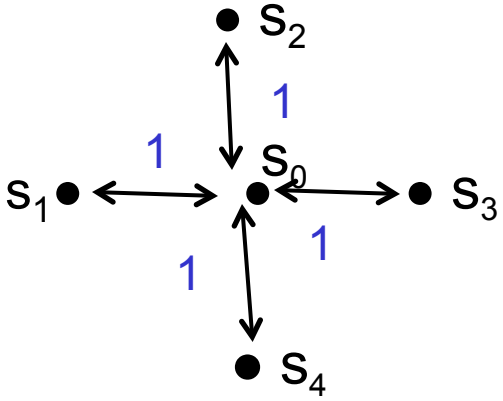


s_2

$$\frac{1}{4}\left[dist(s_2,s_0)+d(s_2,s_1)+d(s_2,s_3)+d(s_2,s_4)\right]=\frac{1}{4}(1+\sqrt{2}+\sqrt{2}+2)=1,45$$

Poblacions centrals o de referència (3)
Quina és la S_{ref}?

En canvi:



$s_0 = s_{ref}$

$$\frac{1}{4}\left[dist(s_0,s_1)+d(s_0,s_2)+d(s_0,s_3)+d(s_0,s_4)\right]=\frac{1}{4}(1+1+1+1)=1$$

Observació: $s_0 = s_{ref}$ ja que és la població que presenta les interdistàncies mínimes amb les altres poblacions.

Poblacions centrals o de referència (4)
Quina és la S_{ref}?

Per a un grup **G** format per les poblacions $\mathbf{s_1}$, $\mathbf{s_2}$, $\mathbf{s_3}$, $\mathbf{s_n}$ en què hem definit una distància lingüística $d(s_i,s_j)$, la població central (o de referència) $\mathbf{s_{ref}}$ satisfà la condició:

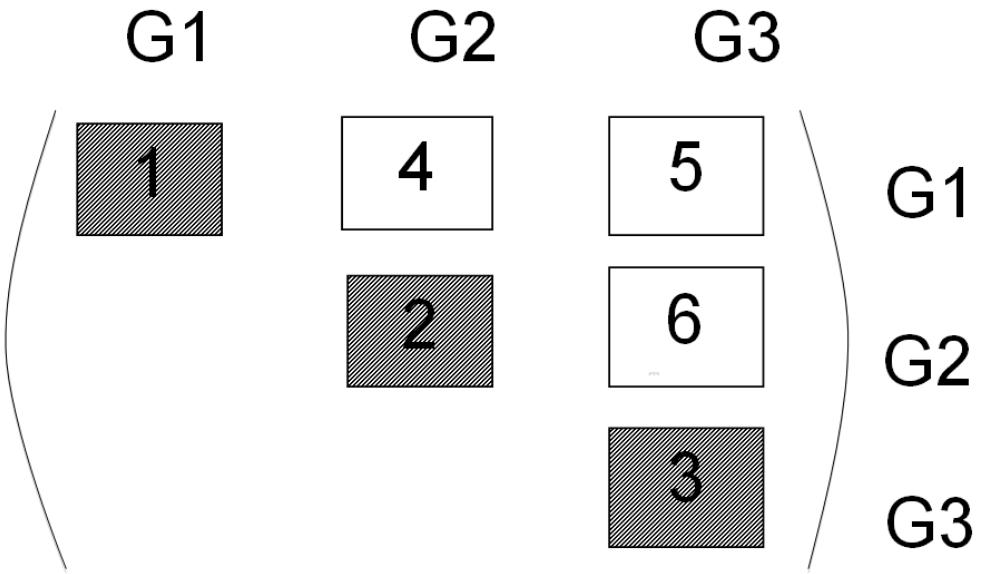
$$dist(s_{ref}, G) = \min_{s_i \in G} dist(s_i, G)$$

on

$$dist(s_i, G) = \frac{1}{n-1} \sum_{s_j \in G} d^2(s_i, s_j)$$

Informativitat (distintivitat) de les formes lingüístiques (1)

Matriu de distàncies per a una forma E1:



$$Inf(E_i) = \frac{\text{interdistància entre grups}}{\text{interdistància intra grups}} = \frac{(4)+(5)+(6)}{(1)+(2)+(3)}$$

Informativitat (distintivitat) de les formes lingüístiques (2)

Seguint aquest esquema, cal buscar les formes més informatives, però sense acumular informació repetida (correlació ≈ 1). Es tracta d'utilitzar arbres de correlació per identificar i descartar els grups que expliquen el mateix.

5. EN CONCLUSIÓ …

Noves estratègies per a l'anàlisi dialectomètrica

- Anàlisi lingüística prèvia, ara a partir de la **teoria de l'optimitat**.
- Anàlisi de la distància lingüística des de la doble perspectiva **espaciotemporal**.
- Anàlisi dialectométrica amb mètodes fuzzy i determinació de les **poblaciones centrals o de referència** i de les **formes més informatives**.

Moltes gràcies per la vostra atenció

Presentació disponible pròximament a: www.ub.edu/GEVAD

Esteve Clua (UPF), esteve.clua@upf.edu
Maria-Rosa Lloret (UB), mrosa.lloret@ub.edu
Miquel Salicrú (UB), msalicru@ub.edu