CILC 2015

VII Congreso Internacional de Lingüística de Corpus

Universidad de Valladolid (5-7 marzo 2015)

El COD2: un corpus oral para el análisis de la variación espacial y temporal del catalán

Esteve Clua
FFI2013-46987-C3-3-P

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

Universitat Pompeu Fabra Barcelona

IULA Institut Universitari de Lingüística Aplicada

UNIVERSITAT DE BARCELONA

Esquema de la presentación

1. ¿Qué es el COD2?
2. Análisis lingüístico
3. Variación espacial + temporal
4. Análisis dialectométrico
5. Conclusión

1. ¿Qué es el COD2?

Corpus Oral Dialectal del catalán contemporáneo (COD2)

- El COD2 es un corpus de las características fónicas y morfológico-flexixas de las variedades geográficas del catalán, actualizado en 2014, basado en un corpus anterior de 1994 (COD).
- La finalidad del corpus es contribuir a ampliar el conocimiento sobre la **variación lingüística** en general y, en particular, sobre la **distancia** entre variedades lingüísticas, desde una doble perspectiva: la **espacial** y la **temporal** (cambio lingüístico entre 1994 y 2014).

COD2 (corpus de 2014)

Proyectos financiados por el MINECO:

- FoCaTeVa (2010-2013, v. www.ub.edu/GEVAD)

  Estudio de la fonología y la morfología del catalán: descripción, teoría y variación (proyecto coordinado UAB, UB y UPF)

- FoMoCaR (2014-2016, v. www.ub.edu/GEVAD)

  Estudio de la fonología y la morfología del catalán y otras lenguas románicas: descripción, teoría y variación (proyecto coordinado UB, UAB y UPF)

- **FoMoCaR** (2014-2016): Estudio de la fonología y la morfología del catalán y otras lenguas románicas: descripción, teoría y variación (proyecto coordinado UB, UAB y UPF)

- ✓ Suproyecto 1 **DIVaL** (UB): **Descripción** e **interpretación** de la **variación lingüística**: aspectos fónicos y morfológicos del catalán y otras lenguas románicas

- ✓ Subproyecto 3 **ADLET** (UPF): **Análisis** de la **distancia lingüística** en los **ejes espacial y temporal**: aspectos fonológicos y morfológicos del catalán

Proyectos anteriores...

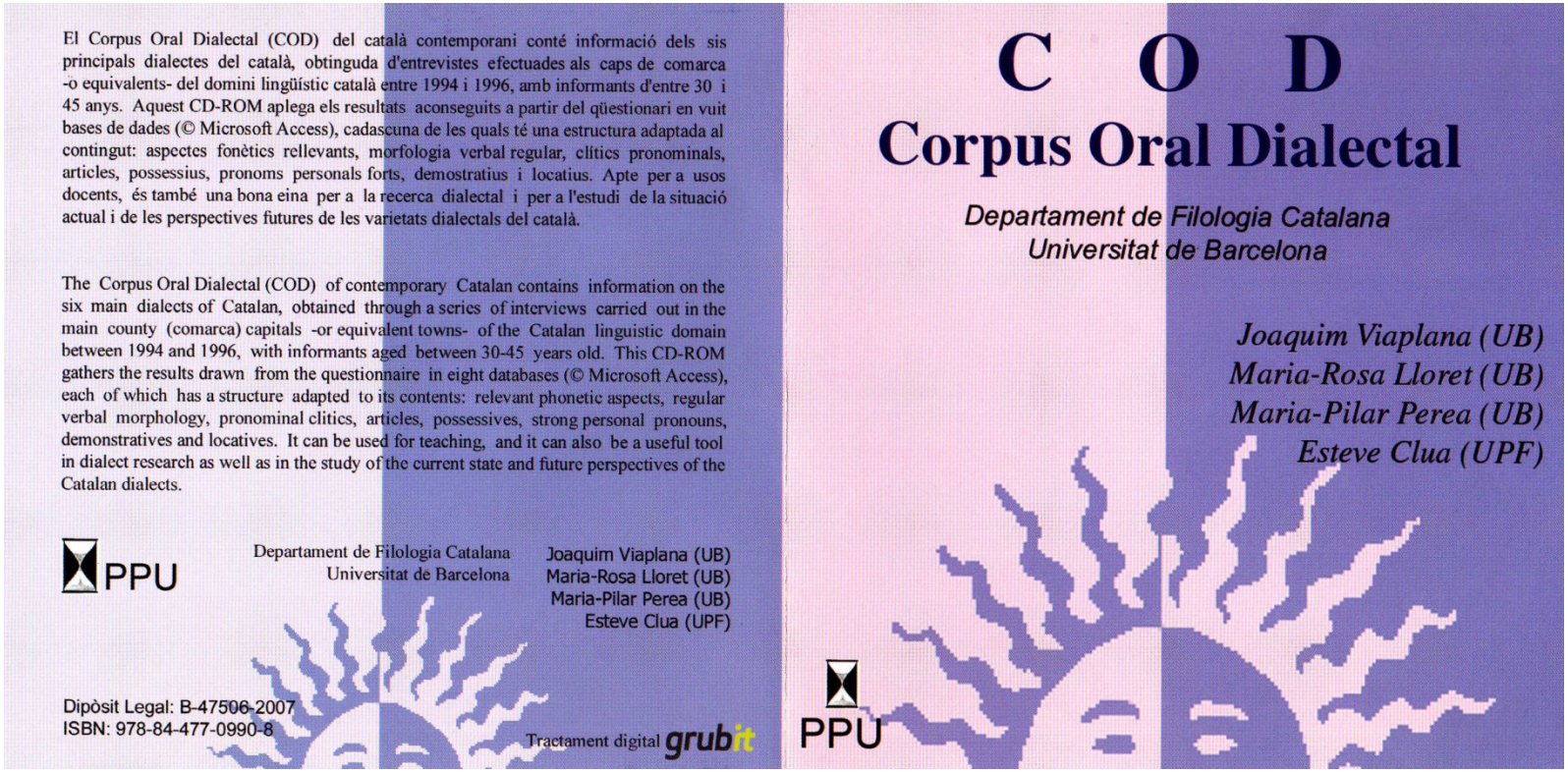
COD (corpus de 1994):

Proyectos financiados por los Ministerios correspondientes :

- .....
- VALDIC (2001-2003)
  Análisis e interpretación de la variación lingüística dialectal a partir de la explotación de un corpus oral
- ECOD, ECOD2 (2004-2010, v. www.ub.edu/lincat)
  Explotación de un corpus oral dialectal: análisis de la variación lingüística y desarrollo de aplicaciones informáticas para la transcripción automatizada

Resultados anteriores (COD)

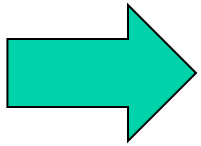
- **http://www.ub.edu/lincat**

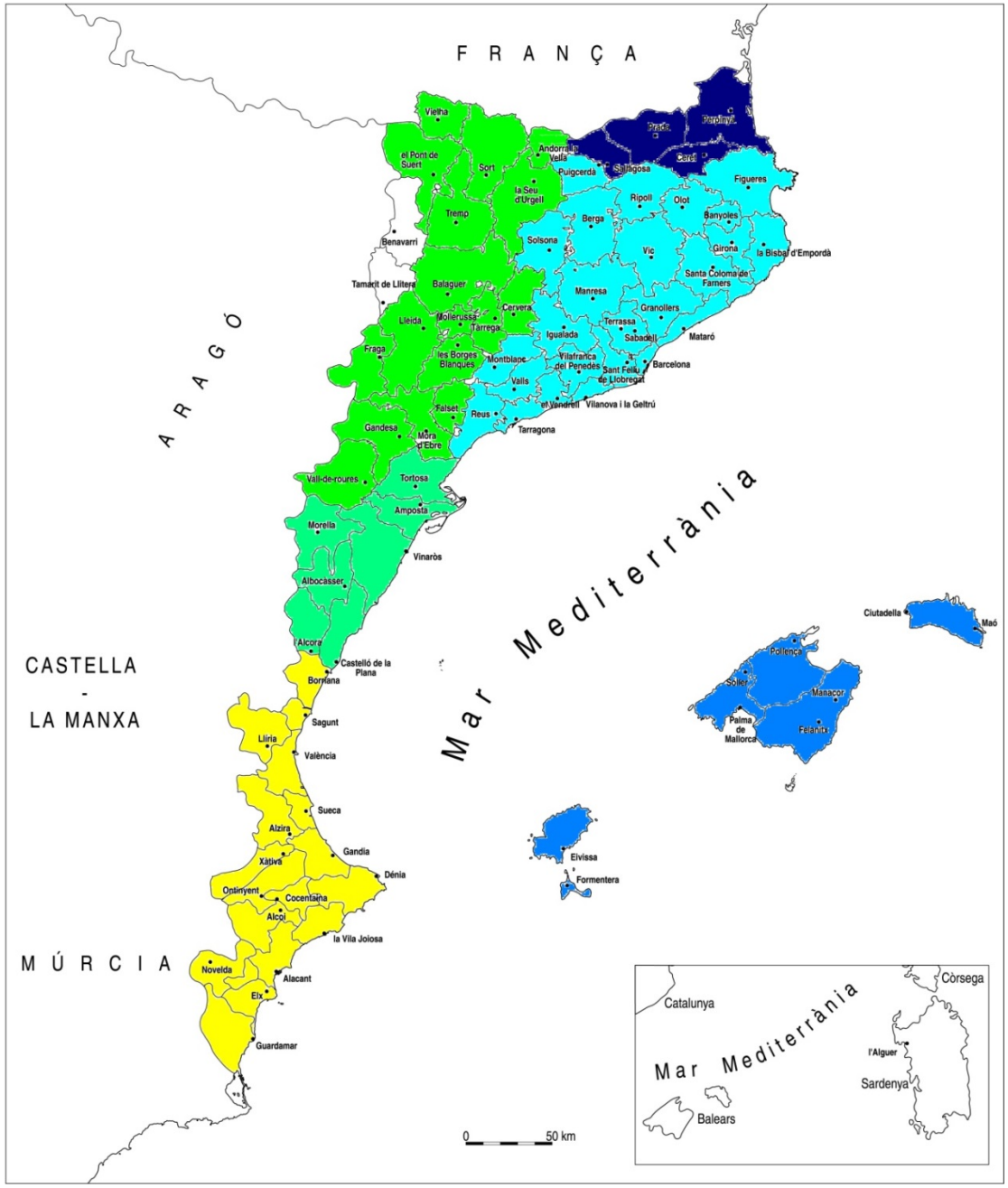


COD2. Características

- Cuestionario de 700 ítems (600 en COD)
- Textos espontáneos (10’)
- 82 capitales de comarca del ámbito lingüístico del catalán
- 3 informantes como mínimo, de 30 a 45 años y de nivel cultural medio

82 capitales de comarca (o equivalentes)





Ámbitos del cuestionario

○ Artículos:	34 preguntas
○ Clíticos pronominales:	78 preguntas
○ Demostrativos:	17 preguntas
○ Posesivos:	22 preguntas
○ Locativos:	3 preguntas
○ Pronombres personales:	8 preguntas
○ Flexión verbal:	250 preguntas
○ Flexión nominal:	20 preguntas
○ Léxico y fonética:	241 preguntas

Conversación libre — 10 minutos

En esta presentación, destacaremos las novedades metodológicas del COD2 en cuanto a:

- Análisis lingüístico (v. § 2)
- Variación espacial + temporal (v. § 3)
- Análisis dialectométrico (v. § 4)

2. Análisis lingüístico

MCOD: Análisis generativo clásico (reglas)

- **Diferencias subyacentes** (computan individualmente):

  Proclítico de 3a pers. masc. sing. acusativo:

  el /l+Ø / - lo /l+o/: 1 diferencia (morfo de masculino)

- **Diferencias superficiales** (computan en forma de **reglas**, de número de cambios):

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

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

COD2: Teoría de la optimidad (restricciones)

- **Diferencias subyacentes** (computan individualmente):

  Proclítico de 3a pers. masc. sing. acusativo:

  el /l+Ø / - lo /l+o/: 1 diferencia (morfo de masculino)

- **Diferencias superficiales** (computan en forma de distancias entre las **restricciones** responsables de las diferencias):

  Proclítico de 1a pers. sing.:

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

Análisis en teoría de la optimidad

- Restricciones que favorecen las vocales menos sonoras (menos prominentes) en posición átona; ordenación universal:

  *ÁTONO/a >> *ÁTONO/e >> *ÁTONO/ə

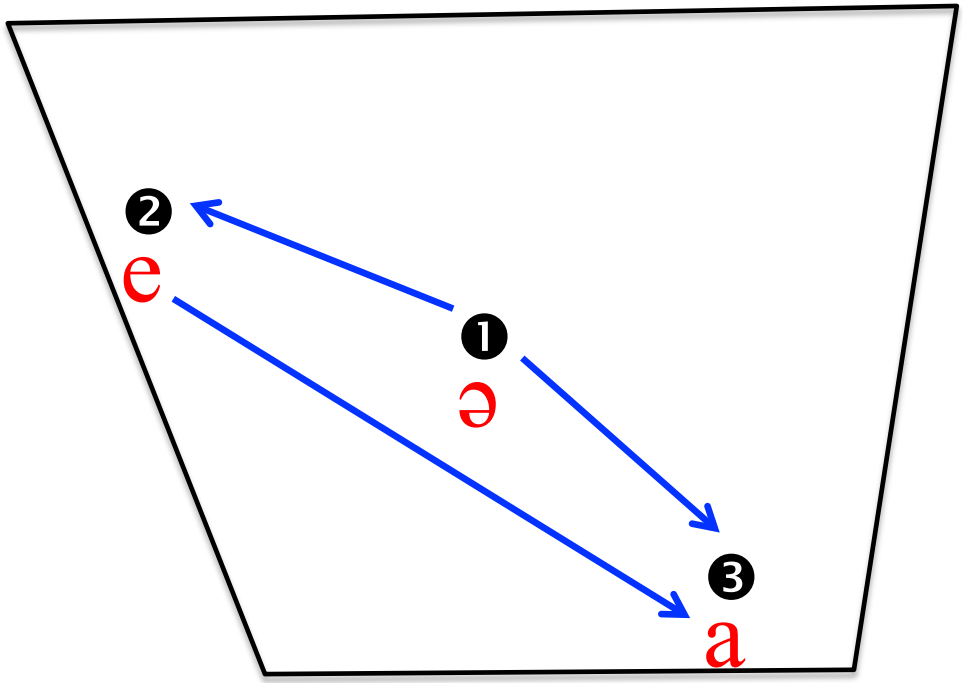
- Restricciones que favorecen, en general y universalmente, que los núcleos silábicos (N) sean más sonoros:

  *N/ə >> *N/e >> *N/a

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

[əm] – [am]: 12 (antes 1 sola diferencia)
[əm] – [em]: 6 (antes 1 sola diferencia)
[em] – [am]: 6 (antes 1 sola diferencia)

[əm] – [am]: 12
[əm] – [em]: 6
[em] – [am]: 6



3. Variación espacial + temporal

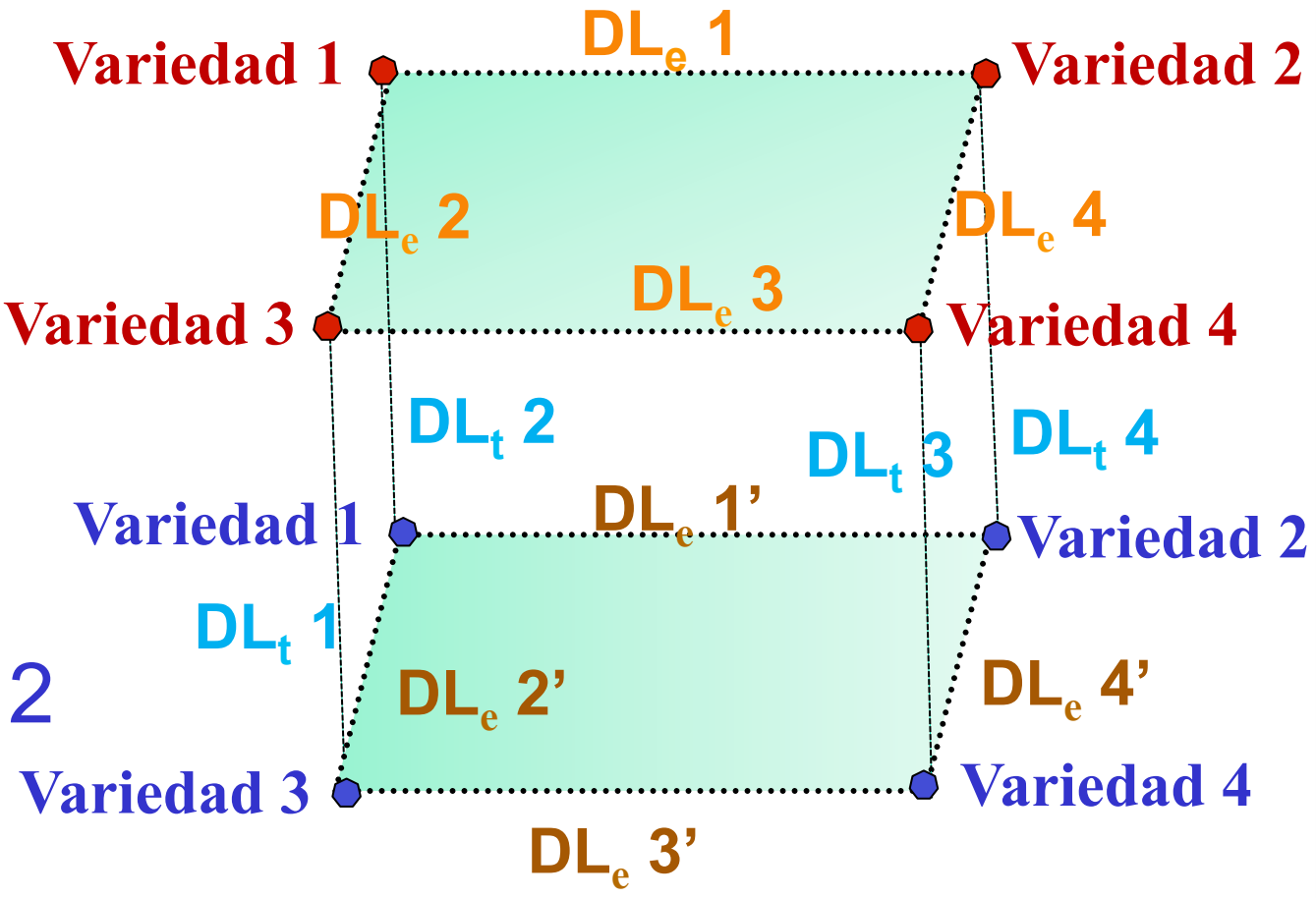
Concepto de Distancia Lingüística (DL)

- Cuantificación de las similitudes o las diferencias lingüísticas entre individuos, poblaciones o grupos de poblaciones

Distancias Lingüísticas: DL_e ESPACIAL DL_t TEMPORAL

1994 COD

2014 COD 2



4. ANÁLISIS DIALECTOMÉTRICO

Métodos dialectométricos

MCOD: método diseñado en los análisis dialectométricos del COD, que se caracteriza principalmente por realizar los análisis cuantitativos a partir de un análisis lingüístico previo de los datos del corpus (antes con reglas, ahora con restricciones).

$$dist(i,j) = \frac{\sum_{k=1}^{long} dif_k(i,j)}{long} \times 100$$

Distancia Levenshtein: es una medida de cálculo de la distancia fonética entre dos líneas de datos. Para determinar esta distancia, el algoritmo de Levenshtein busca el menor conjunto de operaciones básicas necesario para transformar una línea en otra. Estas operaciones pueden ser inserciones, supresiones o sustituciones, y en la versión más simple de la LD tienen las tres un coste de 1.

Herramientas dialectométricas

- **VDM.** Visual Dialectometry de la Escuela dialectométrica de Salzburgo. http://ald.sbg.ac.at/dm/

- **Gabmap**. Aplicación web del Center for Language and Cognition de la Universidad de Groningen (CLCG). http://www.gabmap.nl

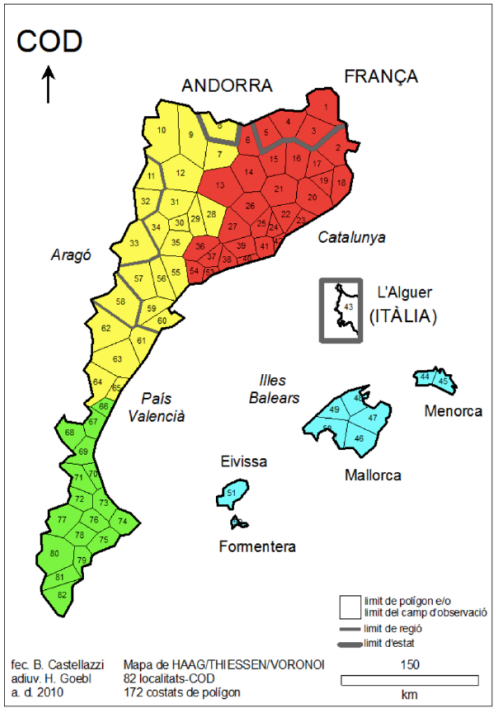
- **DiaTech**. Aplicación web del grupo EUDIA de la Universidad del País Vasco.
  http://eudia.ehu.es/diatech/index/

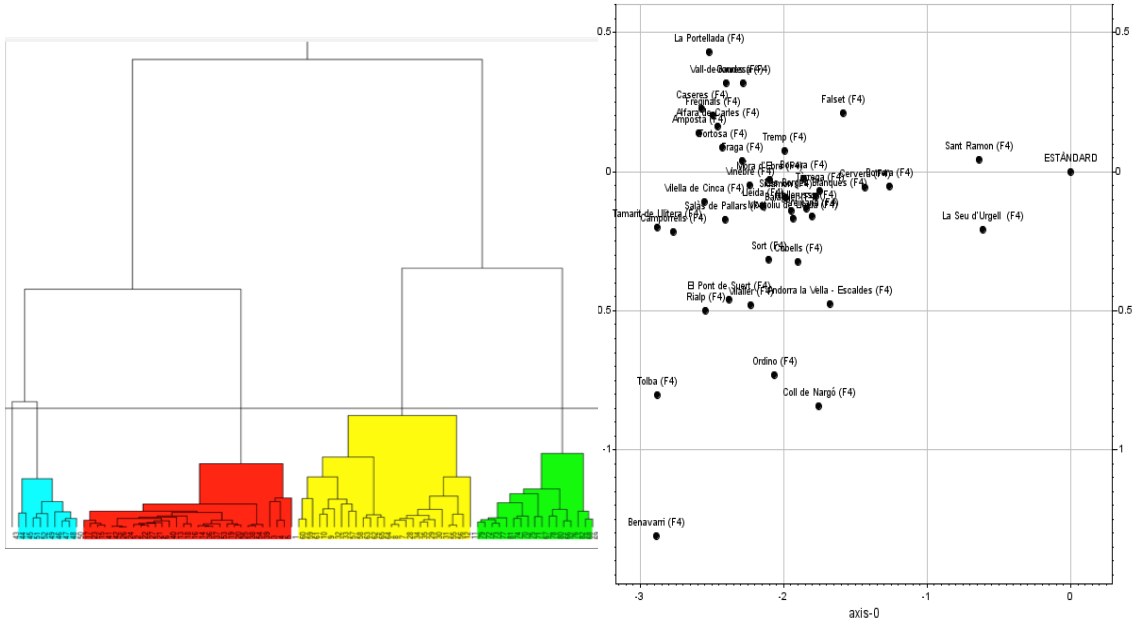
Representación gráfica de la DL

Tipos de representación:

Cartográfica

No cartográfica: dendrográfica, multidimensional

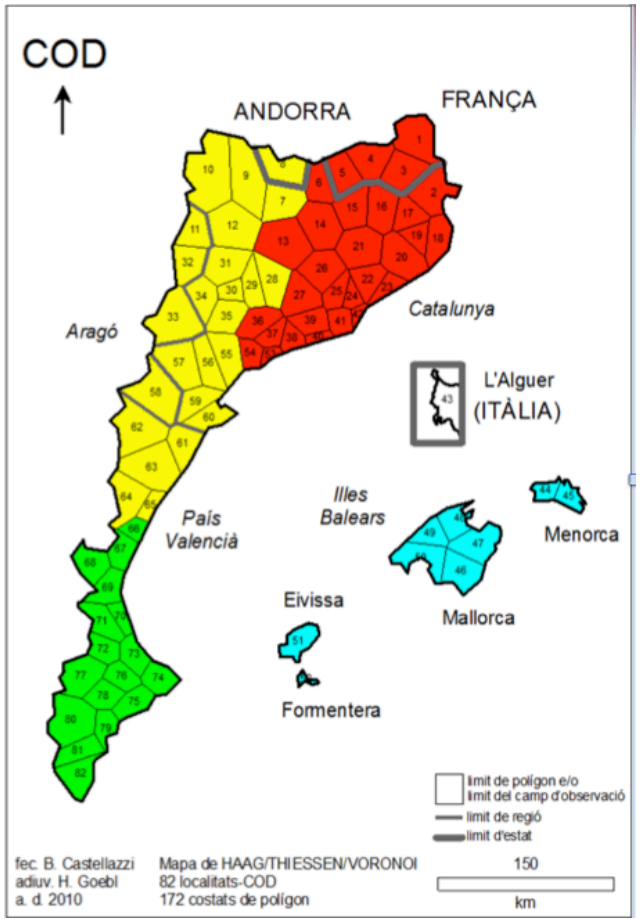




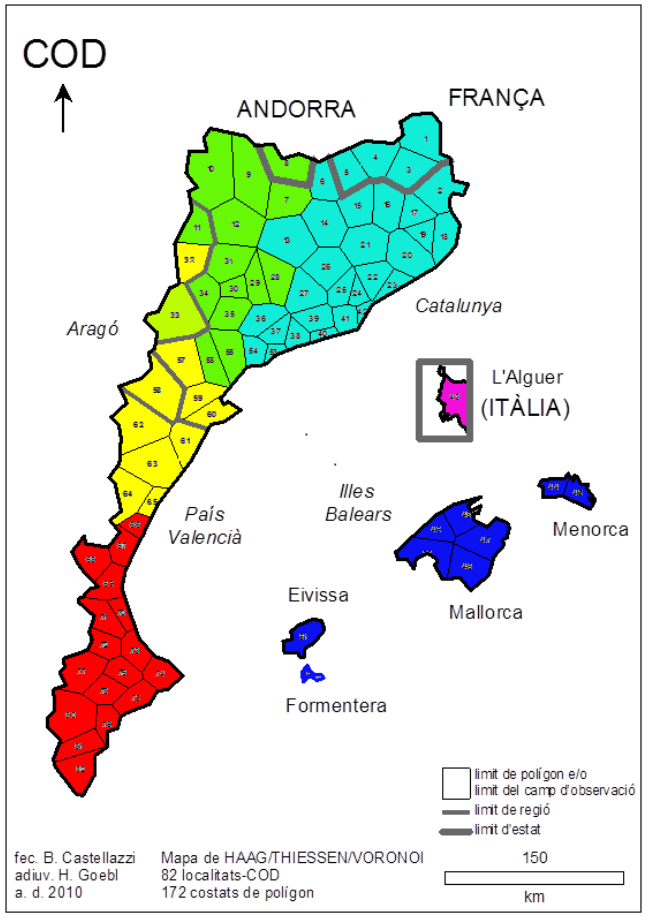
Métodos deterministas vs. métodos fuzzy-probabilísticos

- Los métodos deterministas se orientan a la clasificación de las poblaciones (o variedades) en grupos dialectales, mientras que con métodos fuzzy o probabilísticos la clasificación de las poblaciones se complementa con la verosimilitud de pertenencia a cada grupo. Así, mientras que la clasificación determinista nos conduce a la representación de grupos disjuntos, la representación fuzzy nos proporciona una estructura de grupos más sólida y pone de manifiesto las poblaciones frontera.

Clasificación Determinista



Clasificación FUZZY C-means



Clasificación y caracterización de variedades dialectales

- Uno de los objetivos del análisis dialectométrico del COD2 se centra en identificar las poblaciones y las formas que permiten la caracterización de las variedades lingüísticas que resultan de un proceso de clasificación determinista.

- En cuanto a las poblaciones, el interés se orienta a identificar aquellas que constituyen la referencia de cada grupo (poblaciones centrales o patrón), y en cuanto a las formas, nos interesa identificar las más informativas tanto a nivel global como para cada uno de los grupos; en este último caso se tratará de determinar las formas más representativas y las más distintivas.

5. EN CONCLUSIÓN …

- Análisis lingüístico previo, ahora a partir de la **teoría de la optimidad**.
- Análisis de la distancia lingüística en el eje espacial y ahora también en el eje **temporal**.
- Análisis dialectométrico con métodos probabilísticos y ahora también determinación de las **poblaciones-patrón** y de las **formas más informativas**.

Muchas gracias por su atención

Presentación disponible próximamente en: www.ub.edu/GEVAD

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

Cuestionario

QÜESTIONARI DiLET
2012

Enquestador
25/05/2012

1

1. QÜESTIONARI D'ARTICLES

- Imatge 1

P: Quin és el millor amic de l'home?
R: El gos

P: Quina part del cos es destaca en la imatge?
R: El cap

2

1. QÜESTIONARI D'ARTICLES

P: Quins són els millors amics de l'home?

R: Els gossos

P: Quines parts del cos es veuen a la fotografia?

R: Els peus

1. QÜESTIONARI D'ARTICLES

•Imatge 3

P: On està desada, la roba?

R: A l'armari

P: Qui té els plànols de la casa a la mà?

R: L'arquitecte