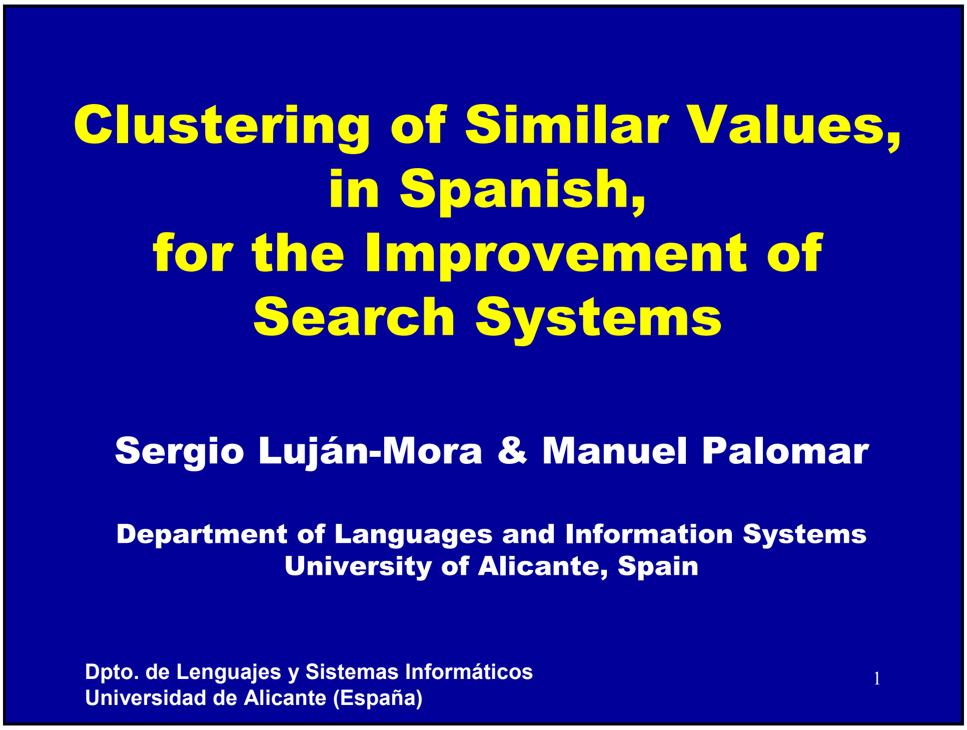
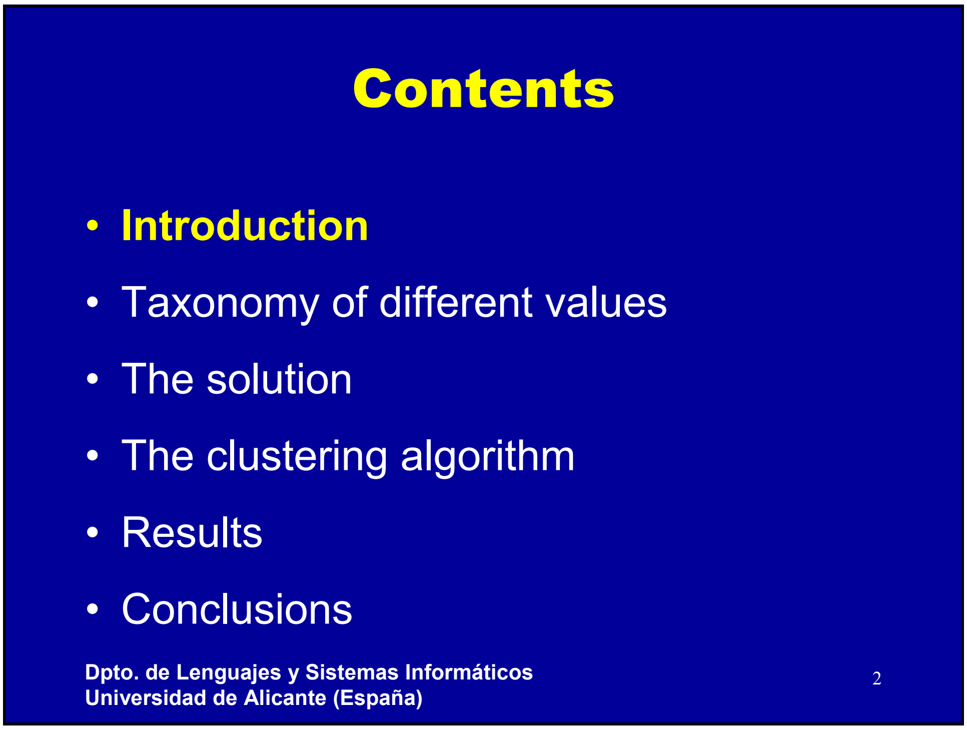
IBERAMIA-SBIA 2000 Open Discussion Track Proceedings, p. 217-226, Atibaia - Sao Paulo (Brasil), November 19-22 2000.



Clustering of Similar Values, in Spanish, for the Improvement of Search Systems

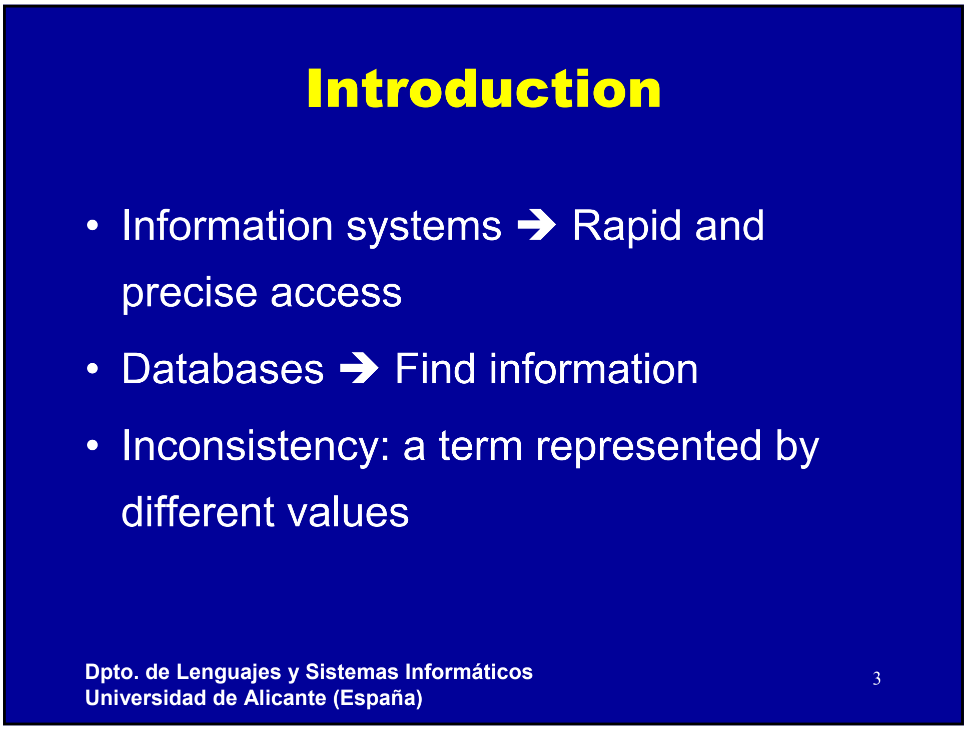
Sergio Luján-Mora & Manuel Palomar

Department of Languages and Information Systems
University of Alicante, Spain



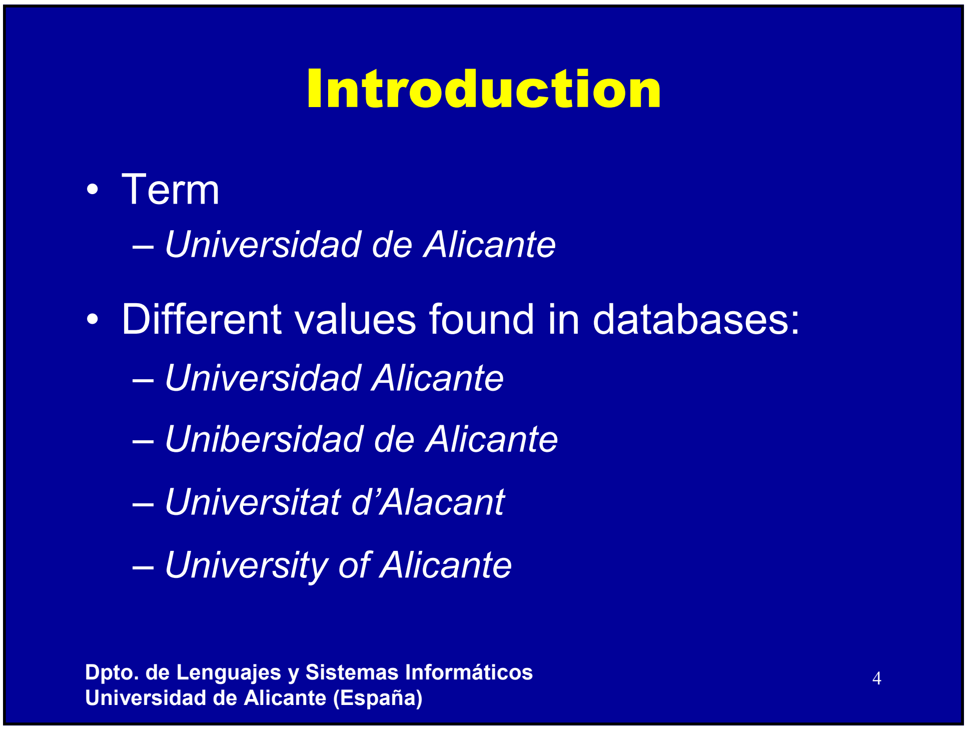
Contents

- **Introduction**
- Taxonomy of different values
- The solution
- The clustering algorithm
- Results
- Conclusions



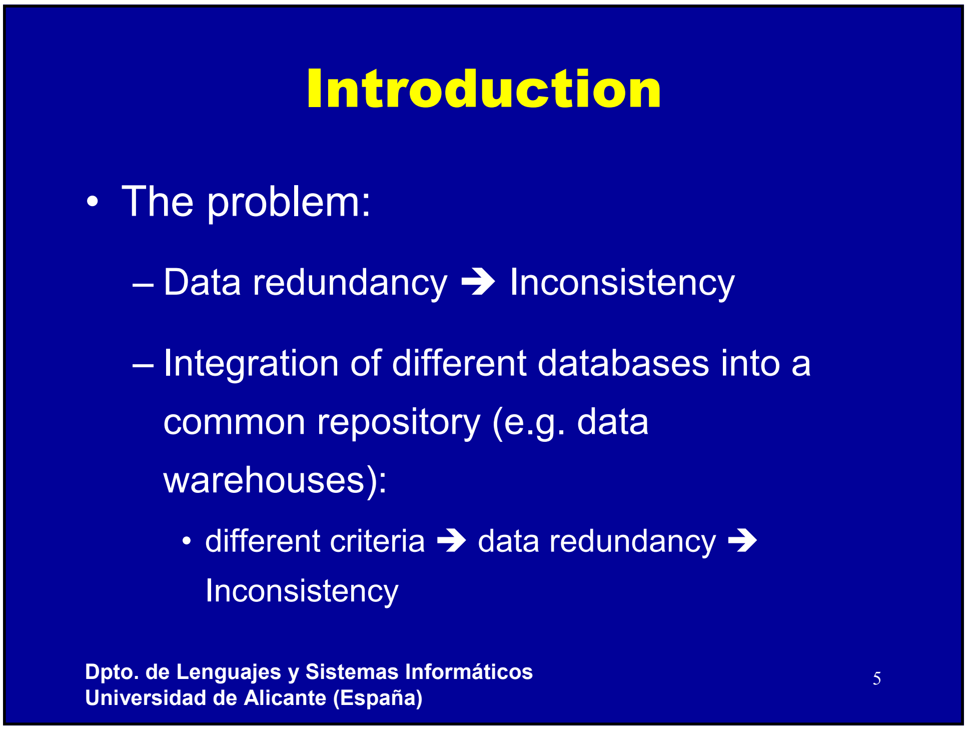
Introduction

- Information systems ➔ Rapid and precise access
- Databases ➔ Find information
- Inconsistency: a term represented by different values



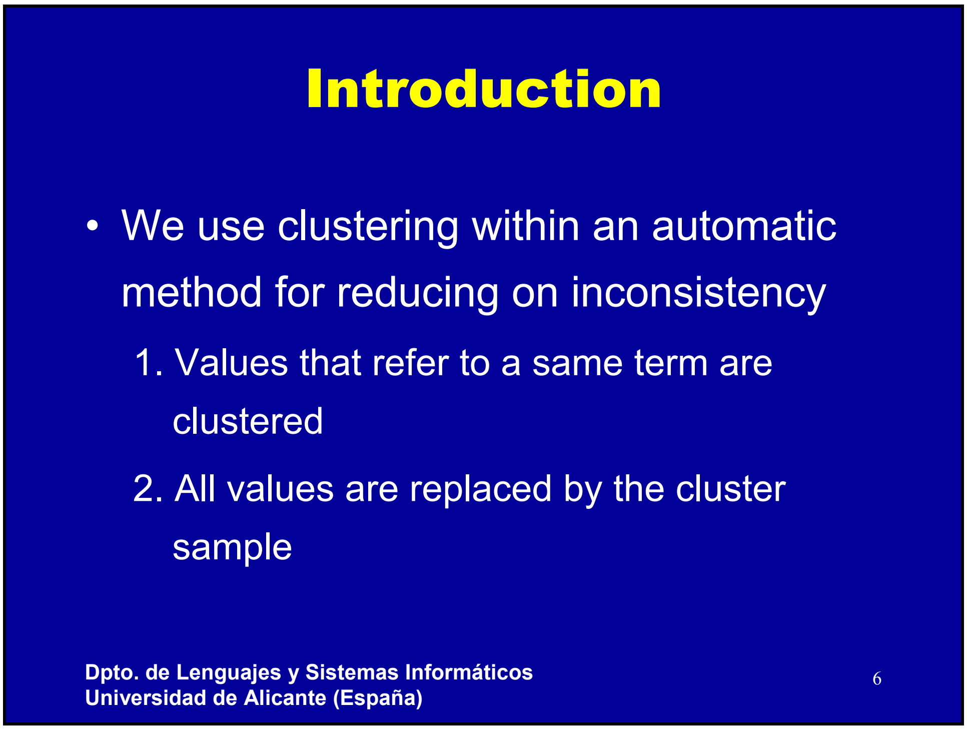
Introduction

- Term
  - *Universidad de Alicante*
- Different values found in databases:
  - *Universidad Alicante*
  - *Unibersidad de Alicante*
  - *Universitat d’Alacant*
  - *University of Alicante*



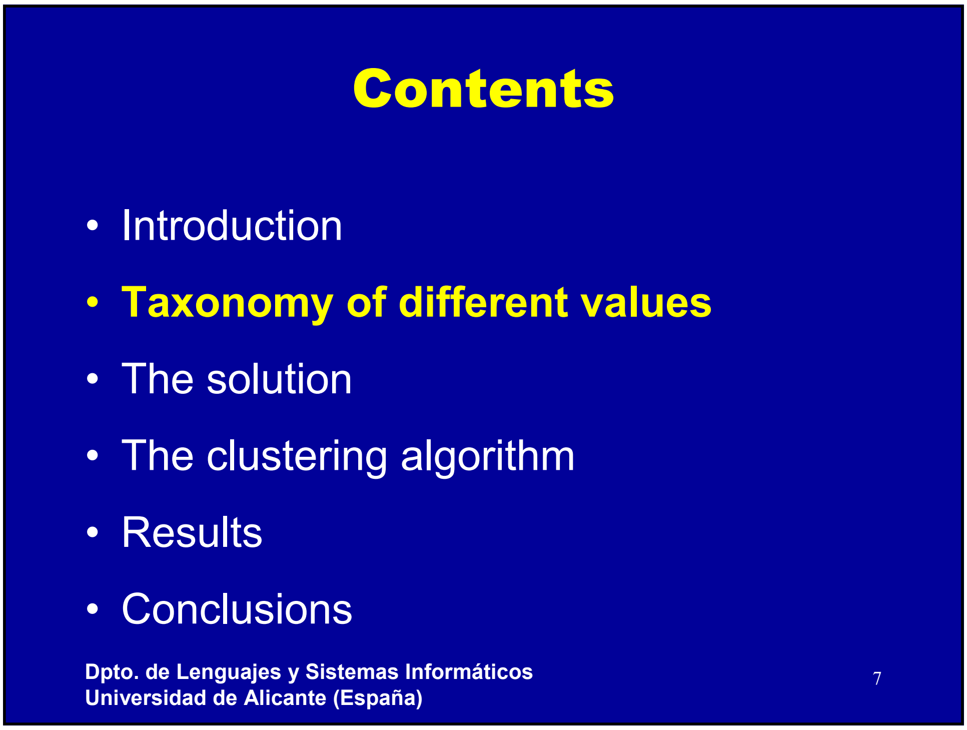
Introduction

- The problem:
  - Data redundancy ➔ Inconsistency
  - Integration of different databases into a common repository (e.g. data warehouses):
    - different criteria ➔ data redundancy ➔ Inconsistency



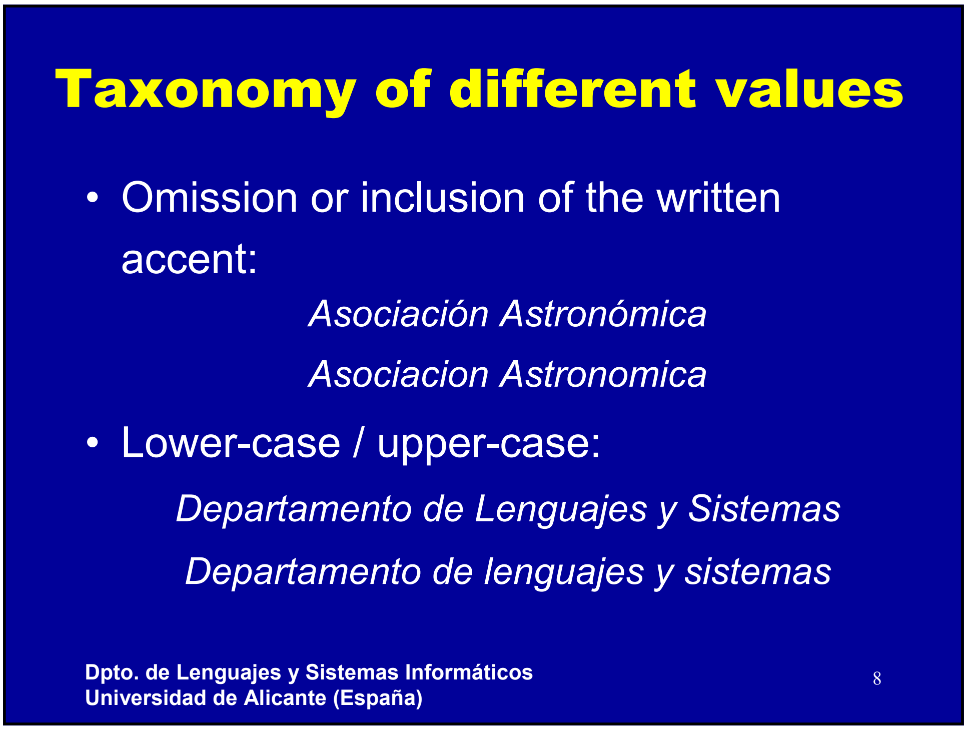
Introduction

- We use clustering within an automatic method for reducing on inconsistency
  1. Values that refer to a same term are clustered
  2. All values are replaced by the cluster sample



Contents

- Introduction
- **Taxonomy of different values**
- The solution
- The clustering algorithm
- Results
- Conclusions



Taxonomy of different values

- Omission or inclusion of the written accent:

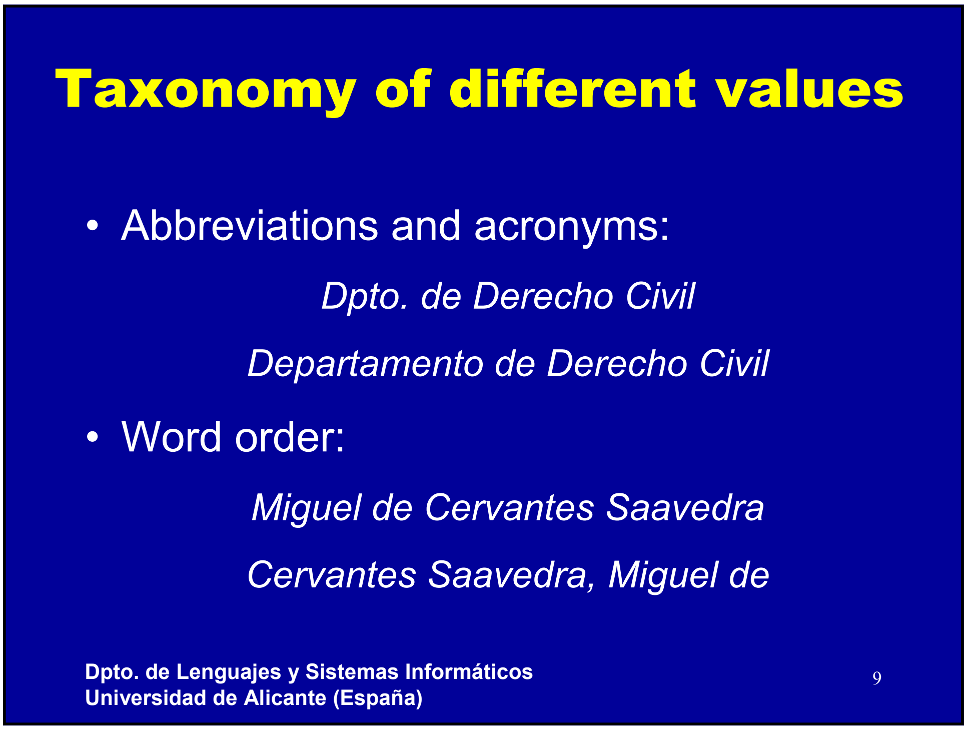
  Asociación Astronómica

  Asociacion Astronomica

- Lower-case / upper-case:

  Departamento de Lenguajes y Sistemas

  Departamento de lenguajes y sistemas



Taxonomy of different values

- Abbreviations and acronyms:

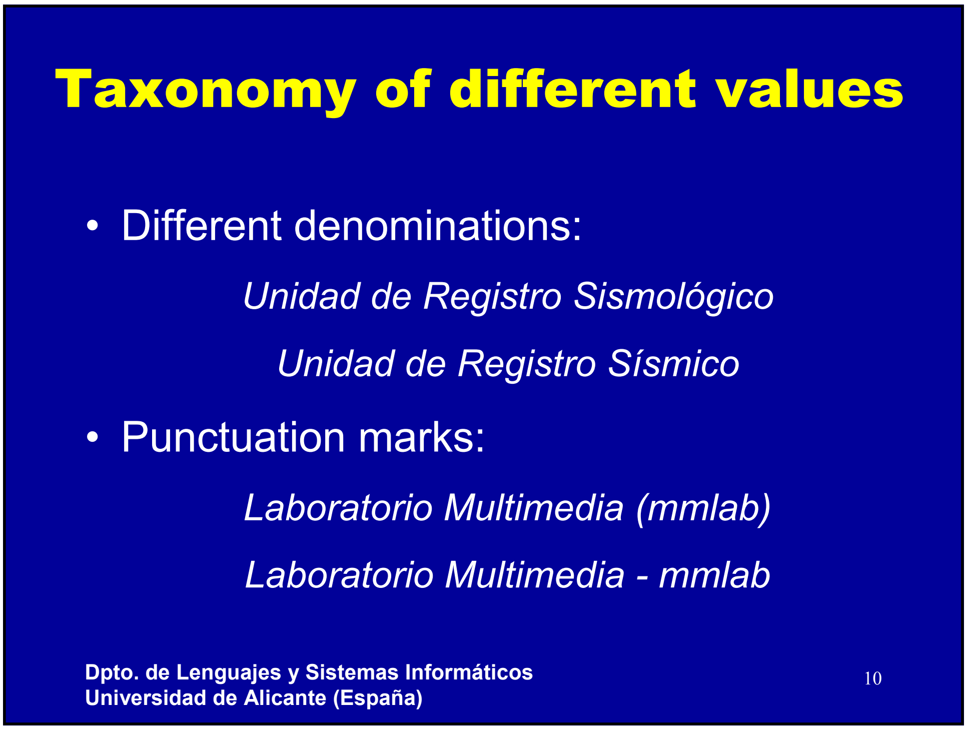
  Dpto. de Derecho Civil

  Departamento de Derecho Civil

- Word order:

  Miguel de Cervantes Saavedra

  Cervantes Saavedra, Miguel de



Taxonomy of different values

- Different denominations:

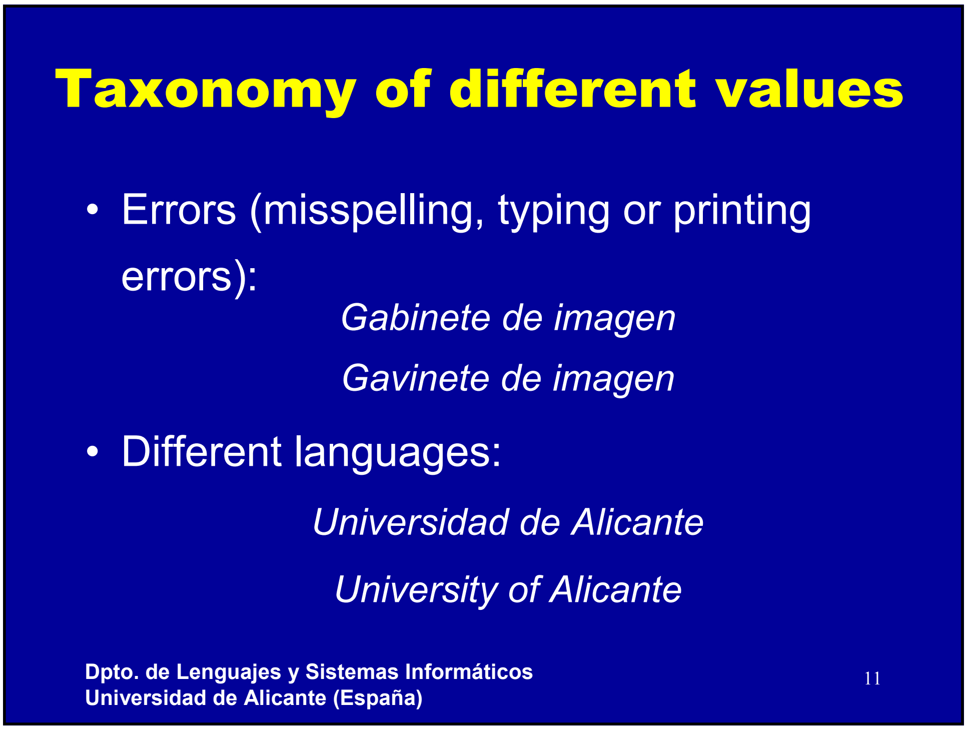
  Unidad de Registro Sismológico

  Unidad de Registro Sísmico

- Punctuation marks:

  Laboratorio Multimedia (mmlab)

  Laboratorio Multimedia - mmlab



Taxonomy of different values

- Errors (misspelling, typing or printing errors):

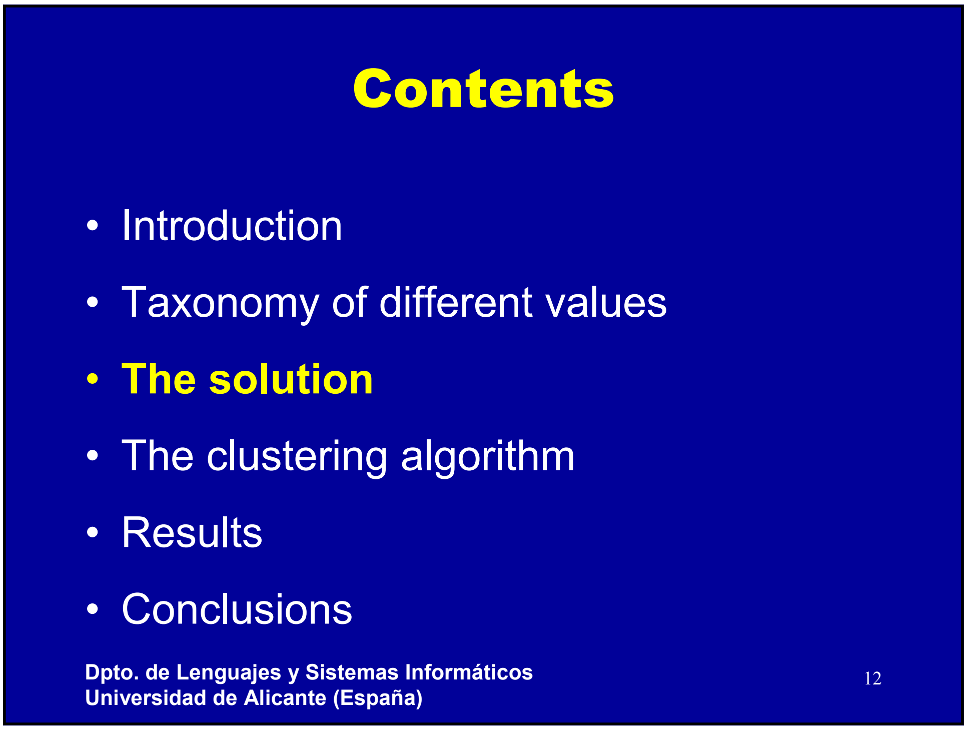
  Gabinete de imagen

  Gavinete de imagen

- Different languages:

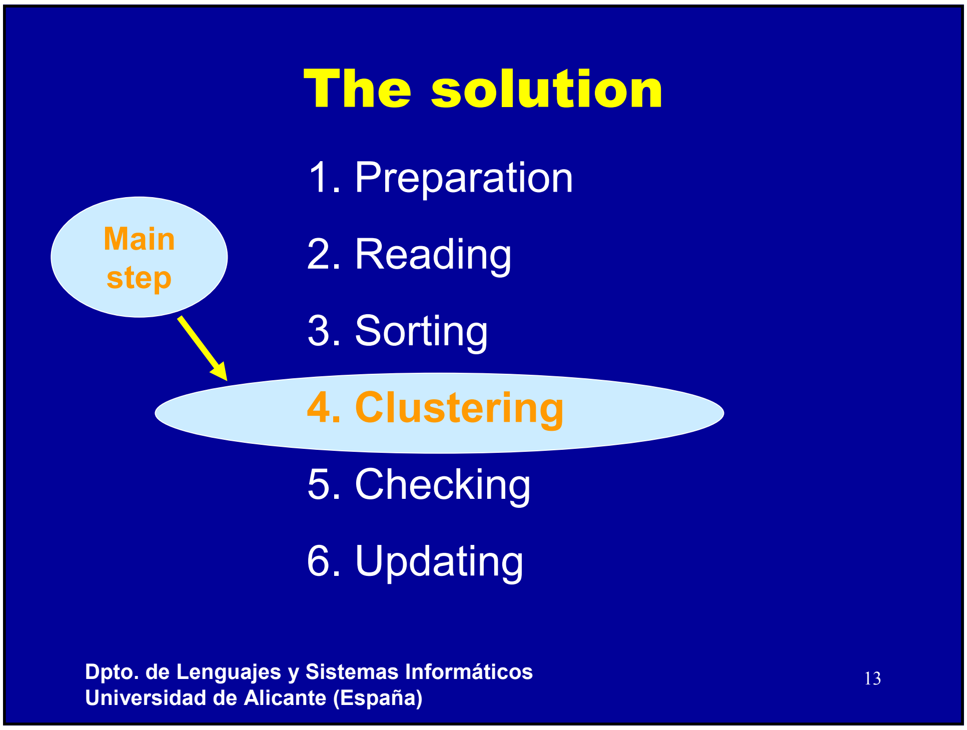
  Universidad de Alicante

  University of Alicante



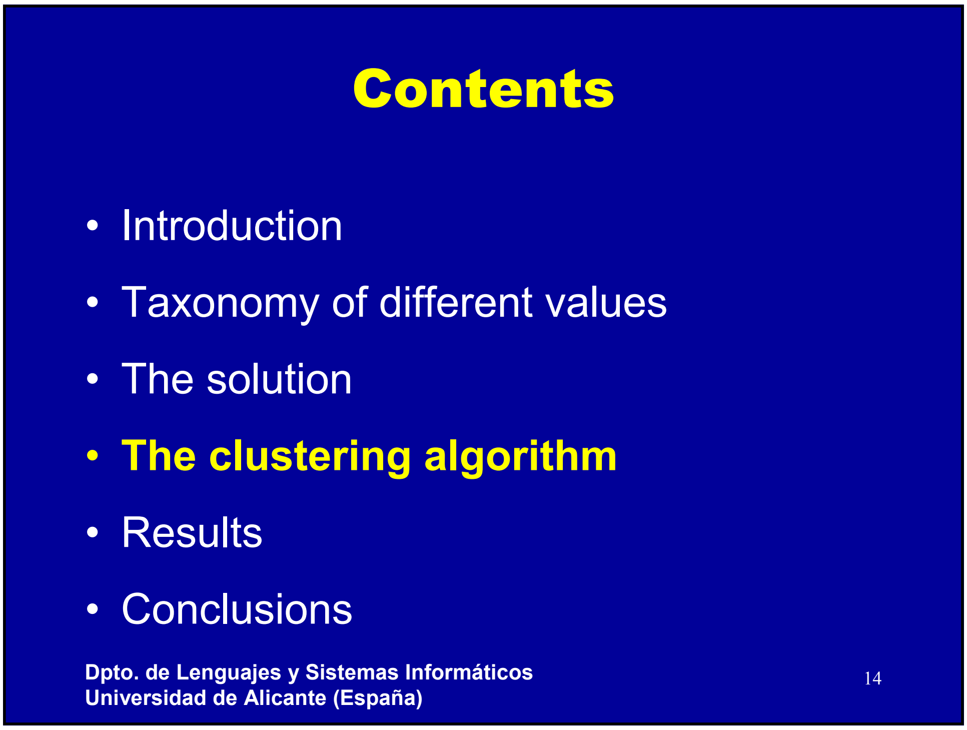
Contents

- Introduction
- Taxonomy of different values
- **The solution**
- The clustering algorithm
- Results
- Conclusions



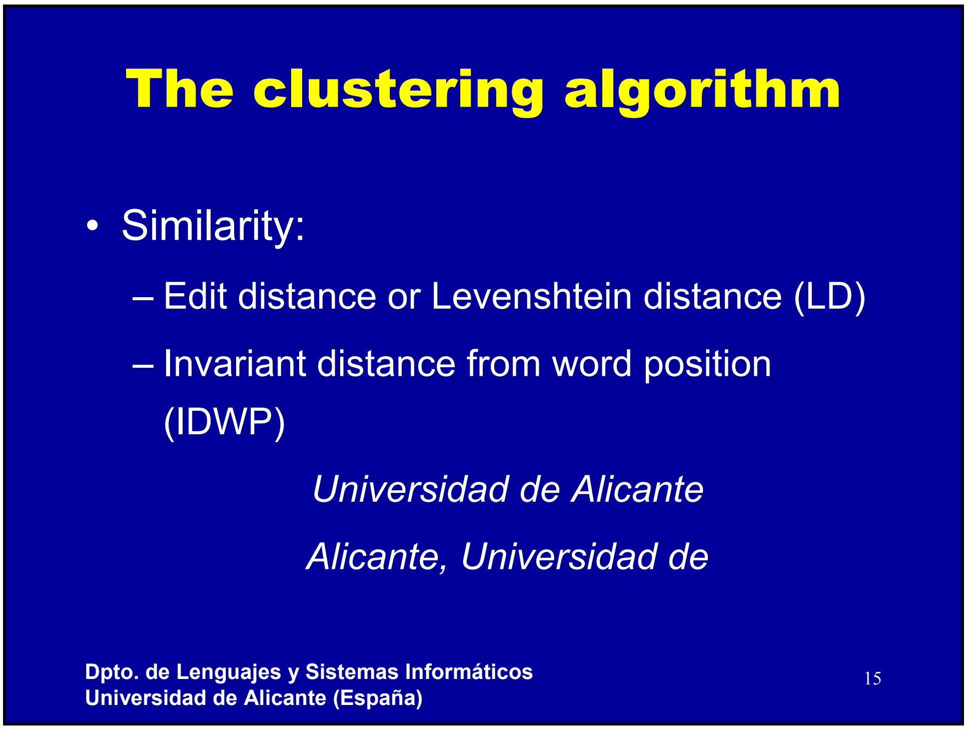
The solution

1. Preparation
2. Reading
3. Sorting
4. **Clustering**
5. Checking
6. Updating



Contents

- Introduction
- Taxonomy of different values
- The solution
- **The clustering algorithm**
- Results
- Conclusions

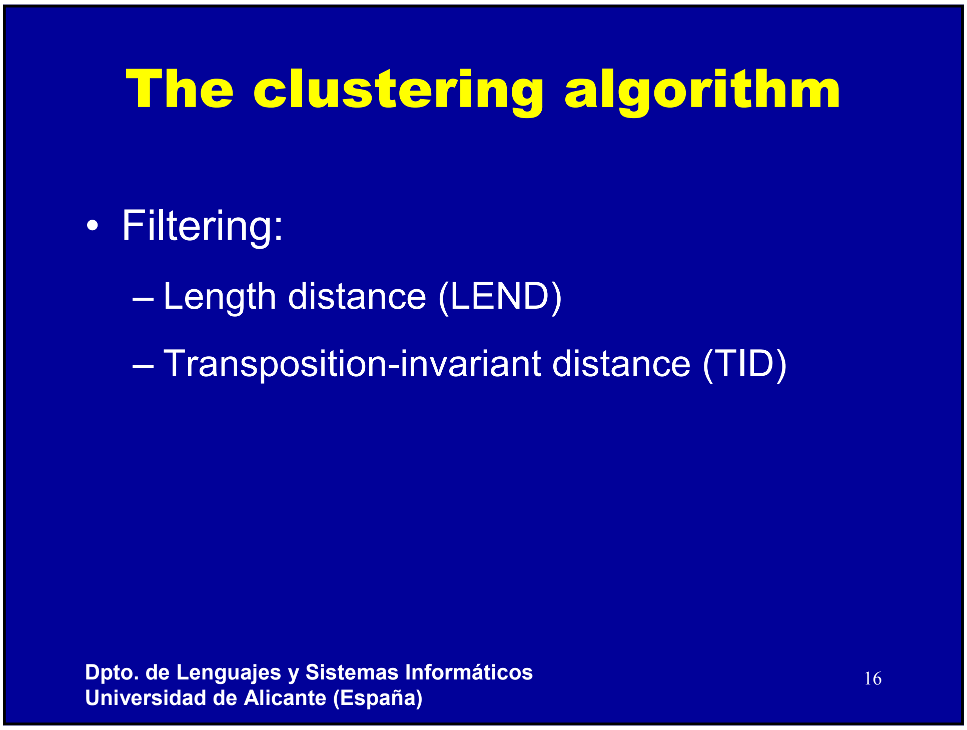


The clustering algorithm

- Similarity:
  - – Edit distance or Levenshtein distance (LD)
  - – Invariant distance from word position (IDWP)

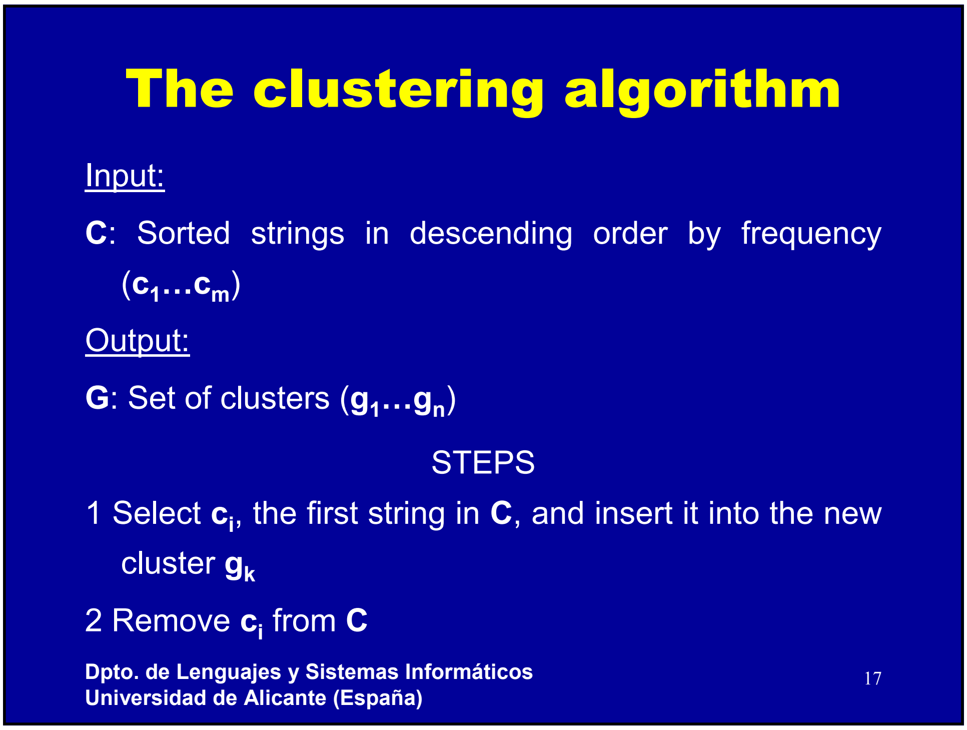
Universidad de Alicante

Alicante, Universidad de



The clustering algorithm

- Filtering:
  - – Length distance (LEND)
  - – Transposition-invariant distance (TID)



The clustering algorithm

Input:

C: Sorted strings in descending order by frequency (**$c_1 \ldots c_m$**)

Output:

G: Set of clusters (**$g_1 \ldots g_n$**)

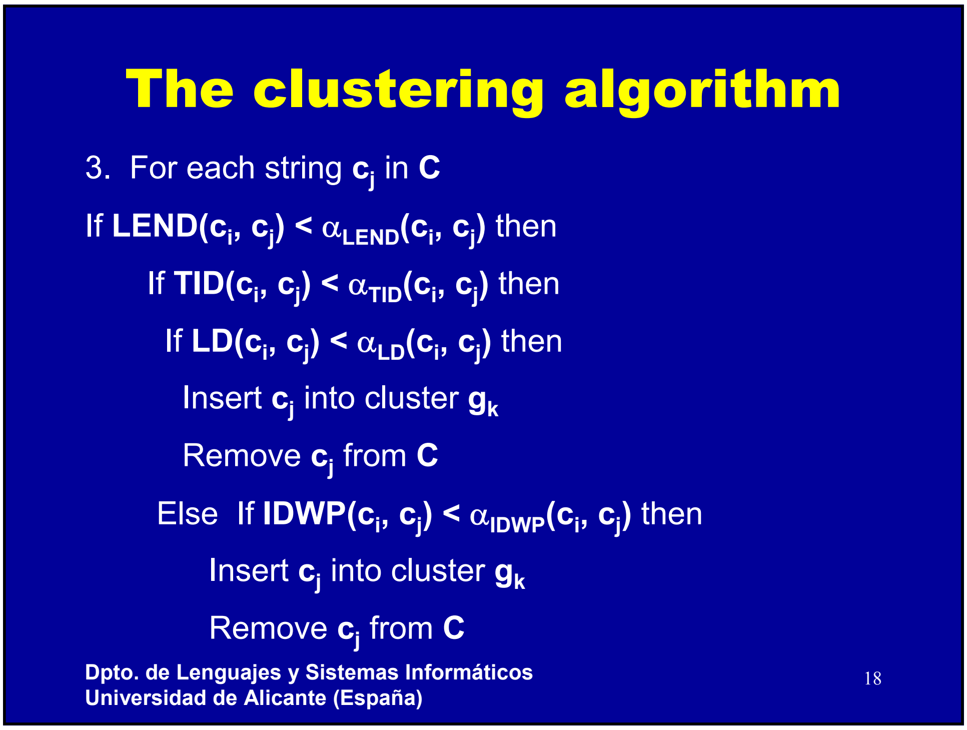
STEPS

1 Select **c_i**, the first string in **C**, and insert it into the new cluster **g_k**

2 Remove **c_i** from **C**

Dpto. de Lenguajes y Sistemas Informáticos
Universidad de Alicante (España)

17

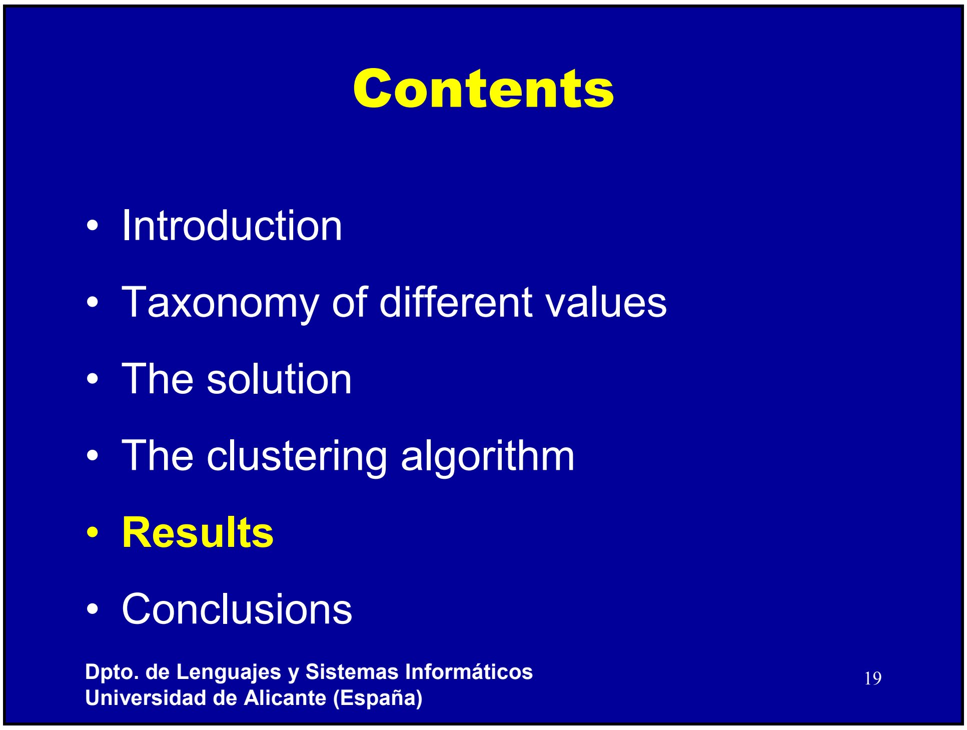


The clustering algorithm

```
3.  For each string $c_j$ in C
If LEND($c_i$, $c_j$) < $\alpha_{LEND}$($c_i$, $c_j$) then
      If TID($c_i$, $c_j$) < $\alpha_{TID}$($c_i$, $c_j$) then
        If LD($c_i$, $c_j$) < $\alpha_{LD}$($c_i$, $c_j$) then
          Insert $c_j$ into cluster $g_k$
          Remove $c_j$ from C
        Else  If IDWP($c_i$, $c_j$) < $\alpha_{IDWP}$($c_i$, $c_j$) then
             Insert $c_j$ into cluster $g_k$
             Remove $c_j$ from C
```

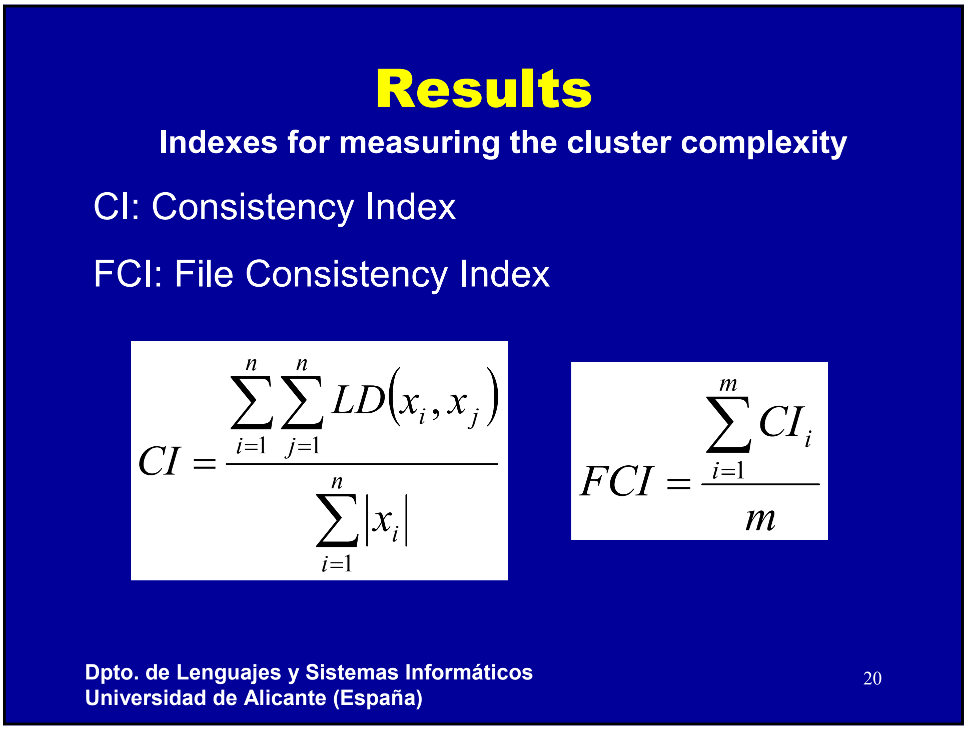
Dpto. de Lenguajes y Sistemas Informáticos
Universidad de Alicante (España)

18



Contents

- Introduction
- Taxonomy of different values
- The solution
- The clustering algorithm
- **Results**
- Conclusions



Results

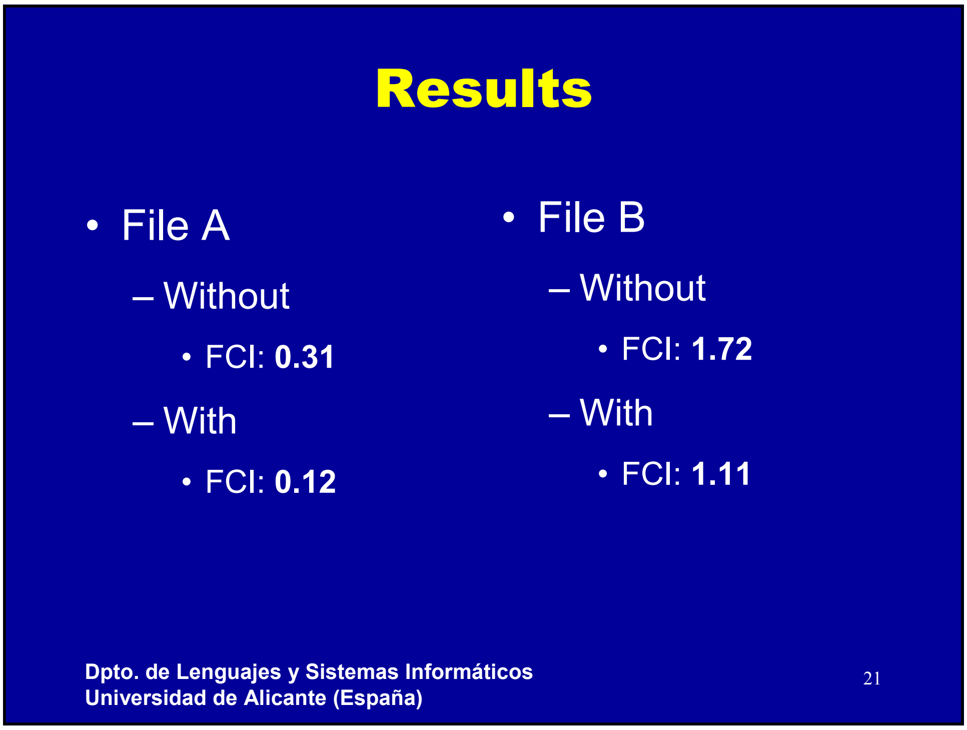
Indexes for measuring the cluster complexity

CI: Consistency Index

FCI: File Consistency Index

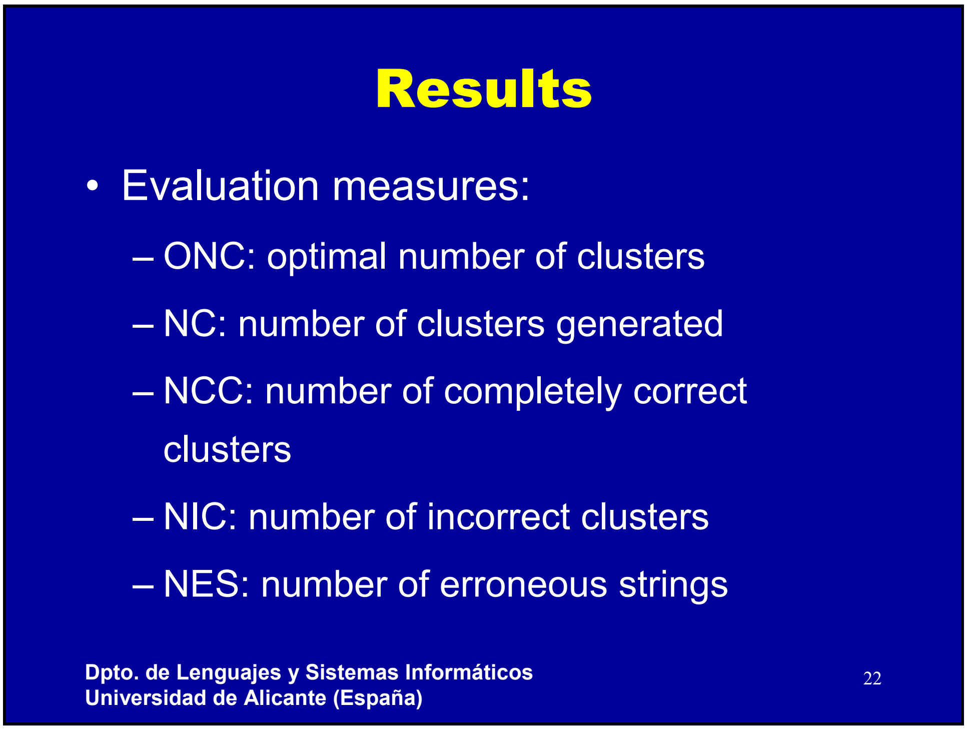
$$CI = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} LD\left(x_i, x_j\right)}{\sum_{i=1}^{n} \left|x_i\right|}$$

$$FCI = \frac{\sum_{i=1}^{m} CI_i}{m}$$



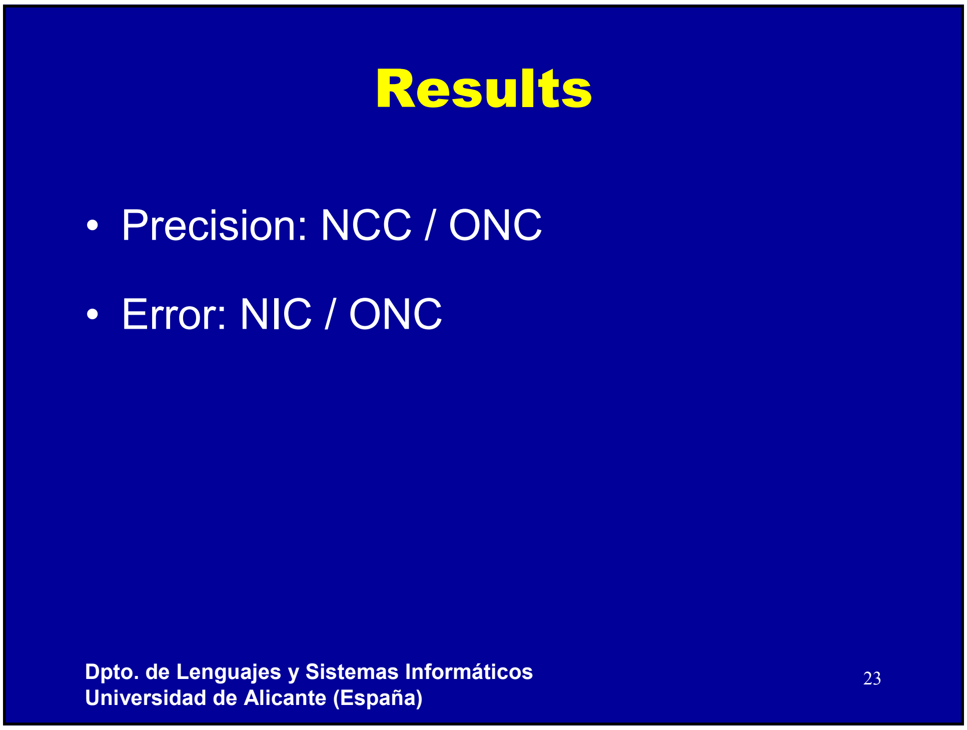
Results

- File A
  - Without
    - FCI: **0.31**
  - With
    - FCI: **0.12**
- File B
  - Without
    - FCI: **1.72**
  - With
    - FCI: **1.11**



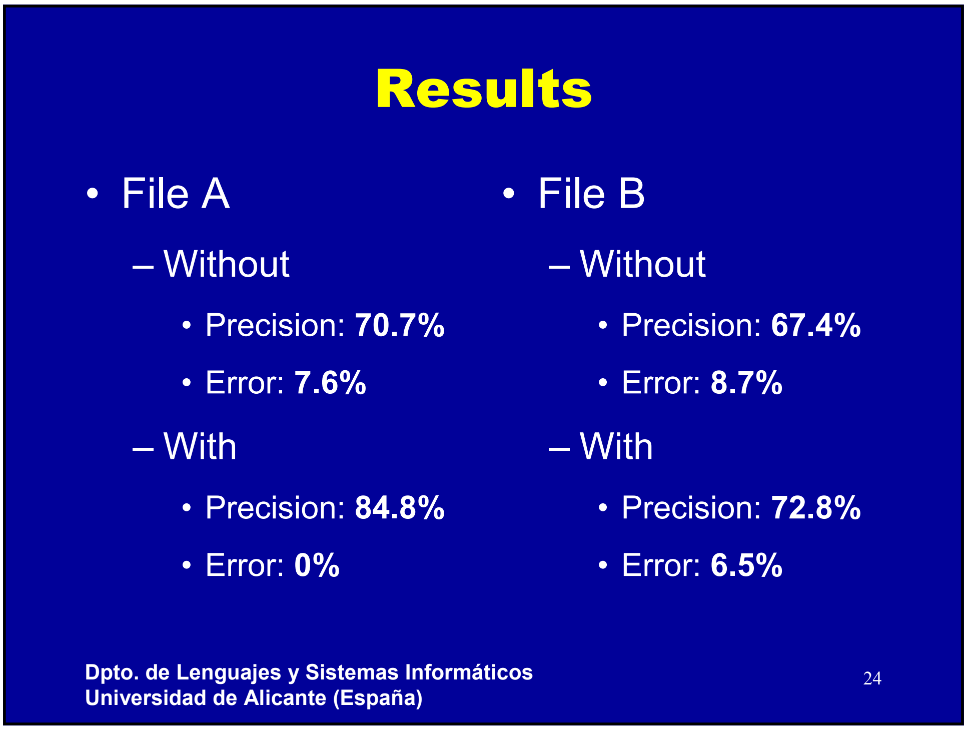
Results

- Evaluation measures:
  - ONC: optimal number of clusters
  - NC: number of clusters generated
  - NCC: number of completely correct clusters
  - NIC: number of incorrect clusters
  - NES: number of erroneous strings



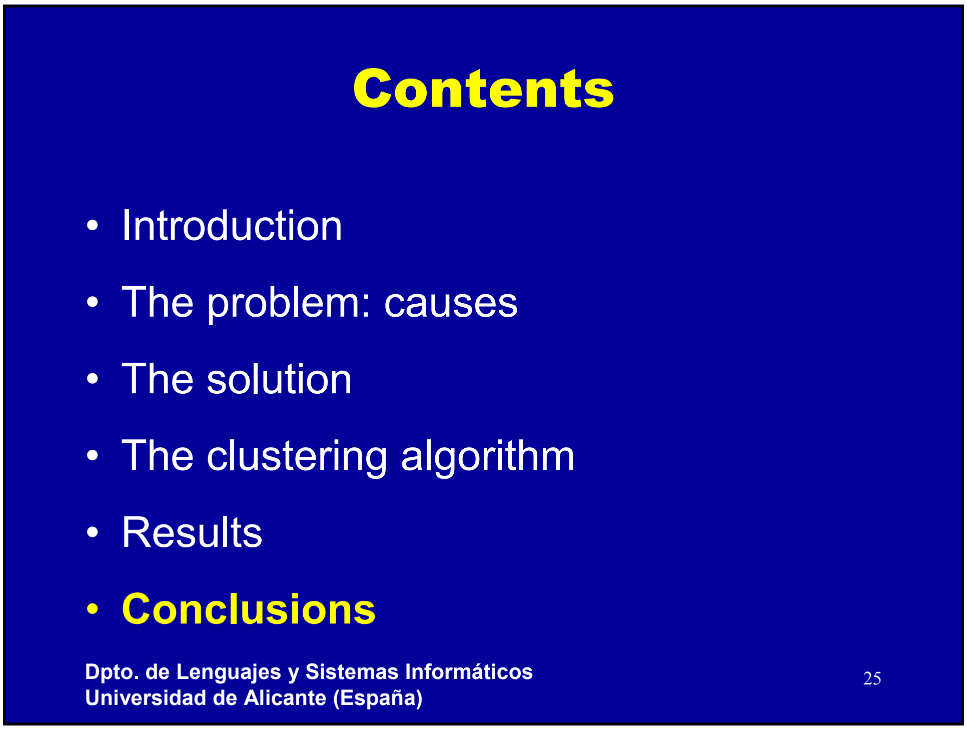
Results

- Precision: NCC / ONC
- Error: NIC / ONC



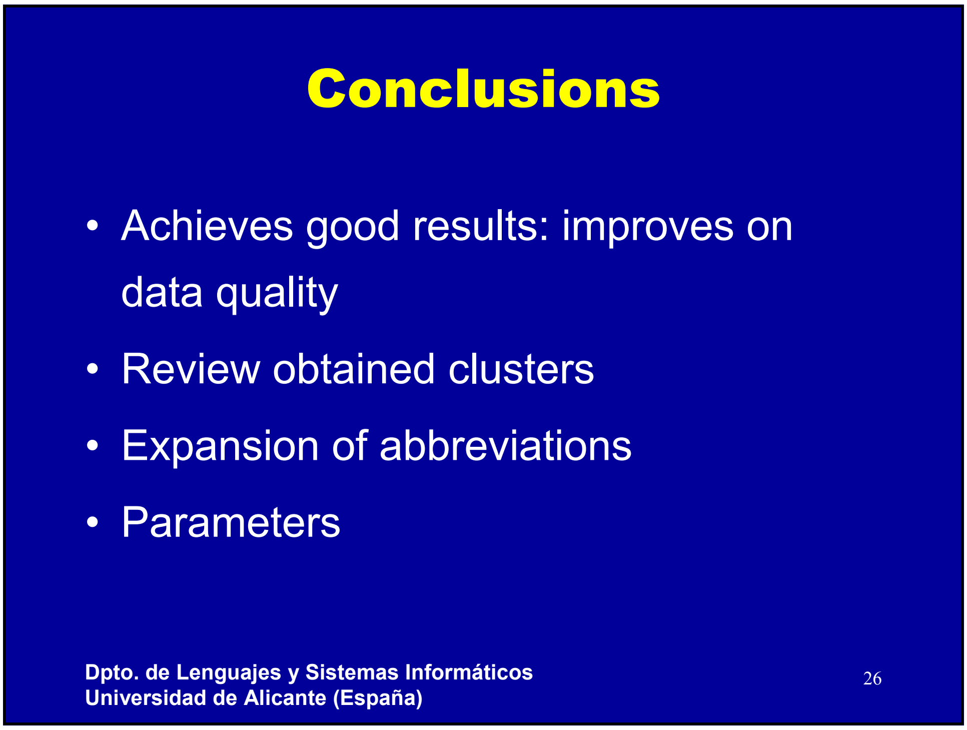
Results

- File A
  - Without
    - Precision: **70.7%**
    - Error: **7.6%**
  - With
    - Precision: **84.8%**
    - Error: **0%**
- File B
  - Without
    - Precision: **67.4%**
    - Error: **8.7%**
  - With
    - Precision: **72.8%**
    - Error: **6.5%**



Contents

- Introduction
- The problem: causes
- The solution
- The clustering algorithm
- Results
- **Conclusions**



Conclusions

- Achieves good results: improves on data quality
- Review obtained clusters
- Expansion of abbreviations
- Parameters