Identificación automática de la temática de un texto

Paz Ferrero García de Jalón

paz.ferrero@uam.es

Universidad Autónoma de Madrid

Universidad Popular de Alcobendas

21-23 de noviembre de 2011

Índice

- Objetivos
- Material de estudio
  - Discursos Navideños del Rey
  - Corpus de referencia (nativos españoles)
  - Exámenes de nivel intermedio de D.E.L.E.
- Métodos de análisis
  - Analizador morfosintáctico FreeLing (lematizador)
  - Campos Semánticos del Plan Curricular del Instituto Cervantes
  - Latent Semantic Analysis
- Conclusiones

Objetivos

Identificación **automática** del contenido de un texto:

- Conocer y cuantificar el contenido de un texto sin leerlo.
- Reconocimiento de textos similares.
- Detección de textos anómalos o fuera de tema.

Material de estudio

- Discursos navideños del Rey
  - Estructura similar
  - Temática común
  - 36 discursos
- Corpus de referencia
  - Textos de nativos españoles con temática semejante a los 40 exámenes de DELE analizados
- Exámenes de D.E.L.E.
  - 40 redacciones sobre 3 temas propuestos en la prueba escrita

Métodos de Análisis

- Campos semánticos: *Glosario* del *PCIC* o *"Índice de nociones generales y nociones específicas"*.
- Latent Semantic Analysis.
- *Corpus* de referencia

Campos Semánticos del *PCIC*

- 7 Nociones generales, y subcampos:

Existenciales, cuantitativas, espaciales, temporales, cualitativas, evaluativa y mentales

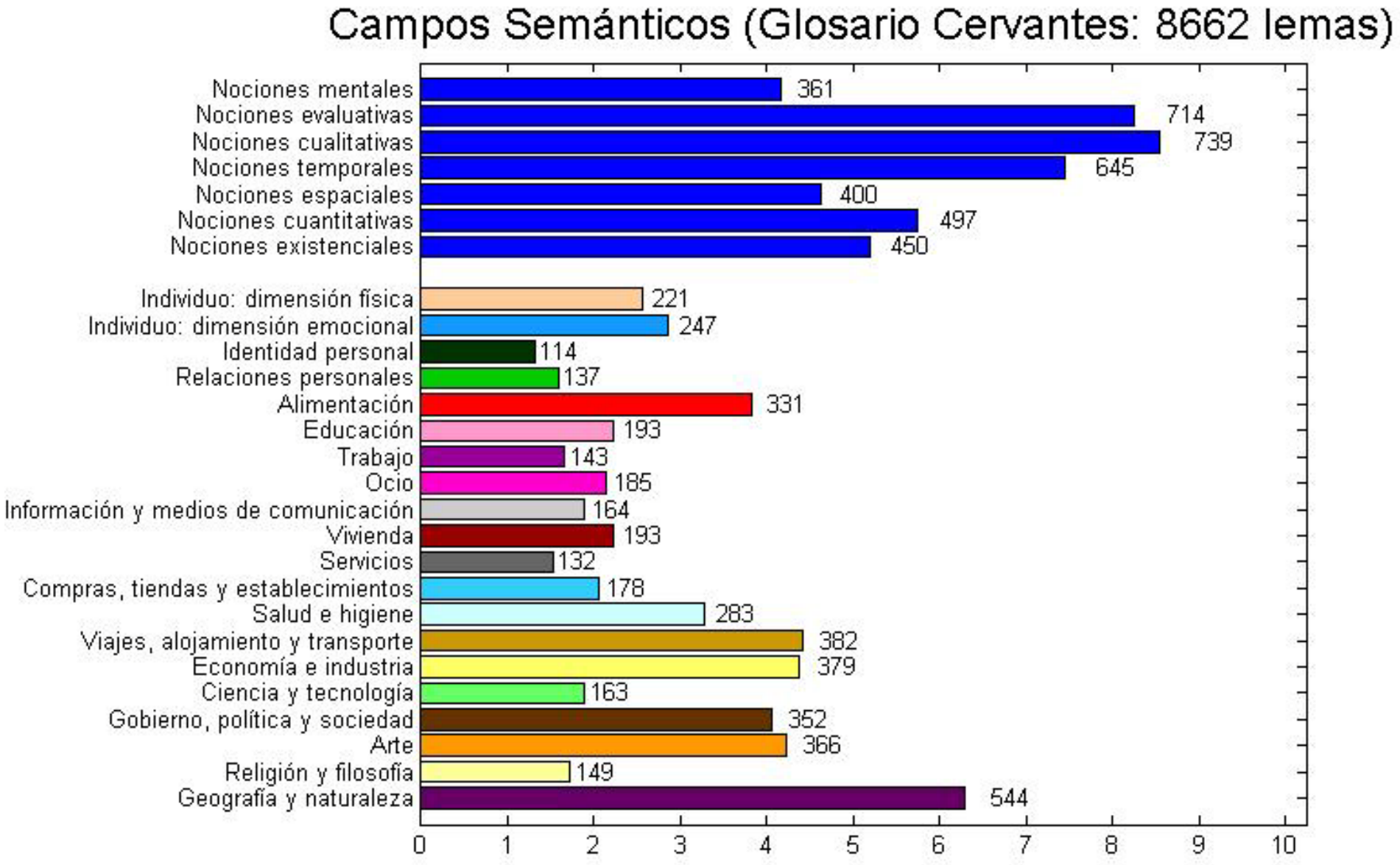
- 20 Nociones específicas y subcampos:

Personas, alimentación, educación, trabajo, ocio, medios de comunicación, vivienda, comercio, salud, viajes, economía, ciencia, política, arte, pensamiento y naturaleza.

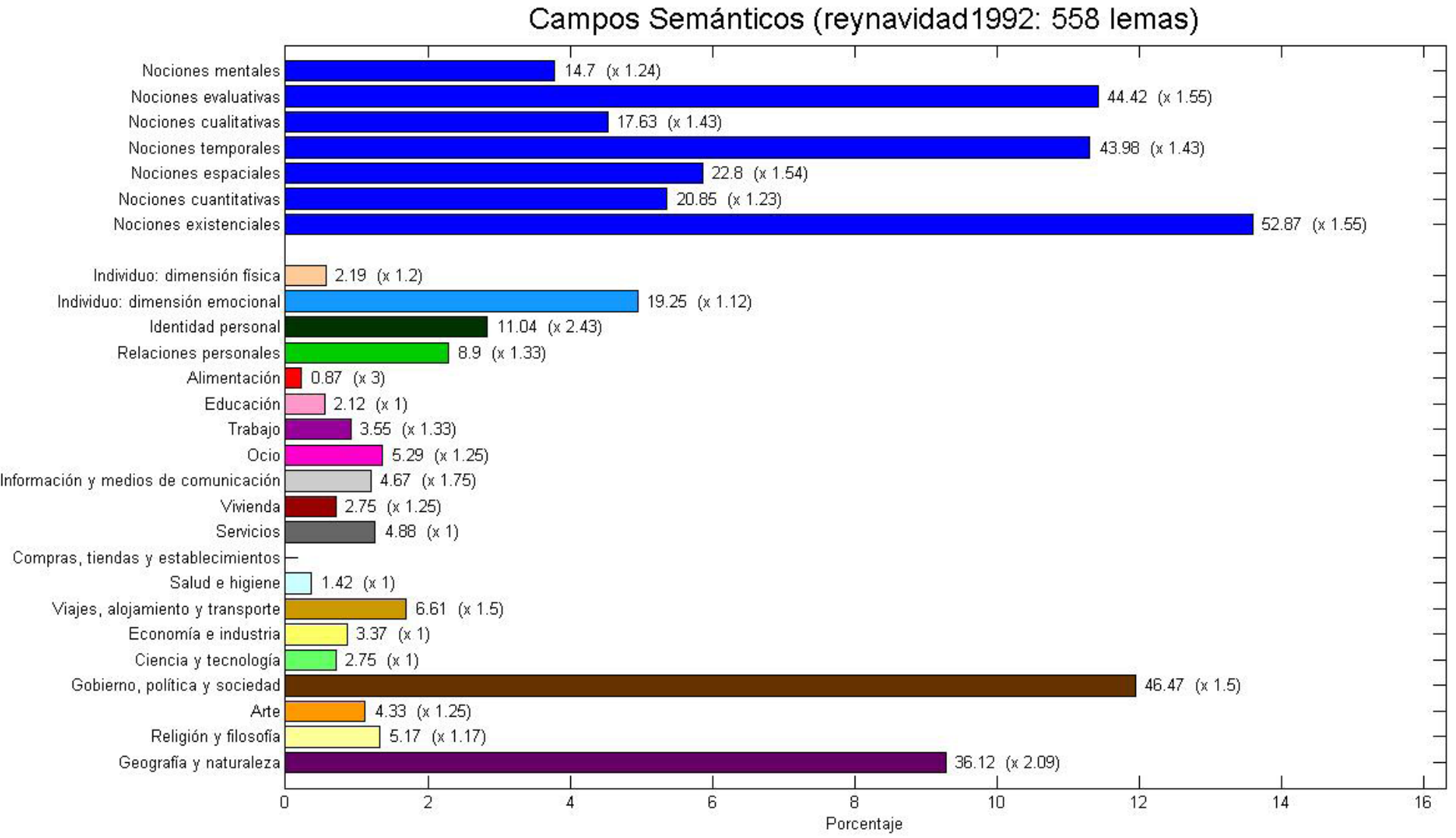
Campos semánticos del PCIC

- Cálculo de las frecuencias de aparición de lemas en un texto en funcion del uso de:
  - Stop List
  - Lemas repetidos
  - Ponderación de ciertos campos en función de los campos semánticos de otros lemas dentro de la frase como desambiguación semántica

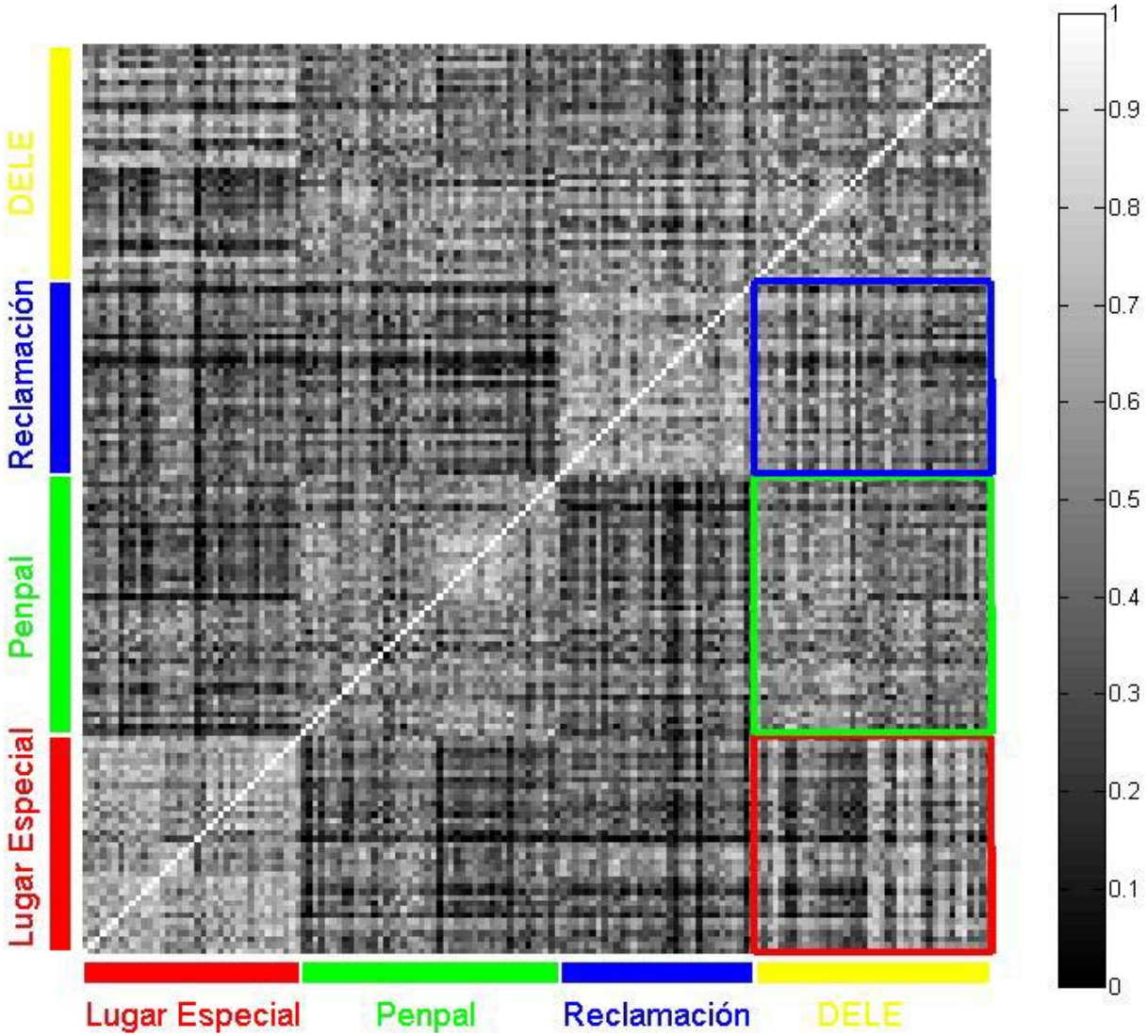
Distribución de lemas del PCIC



Ponderación de campos semánticos

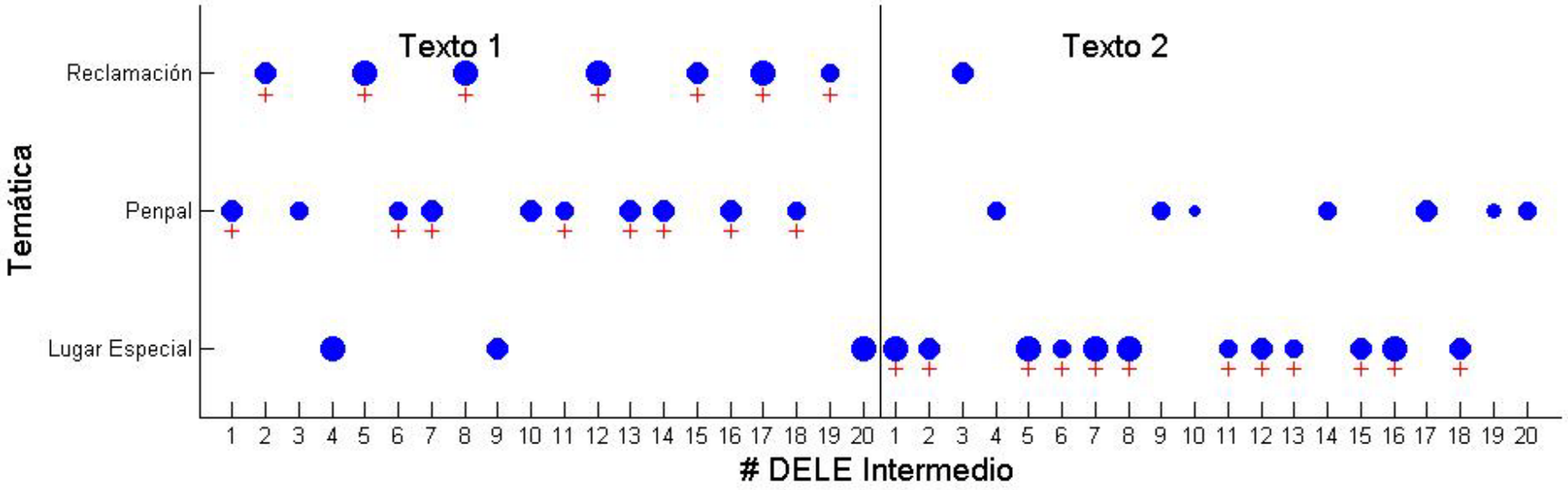


Correlación de textos nativos y DELE



Campos Semánticos del PCIC

Análisis de 40 exámenes de DELE Intermedio



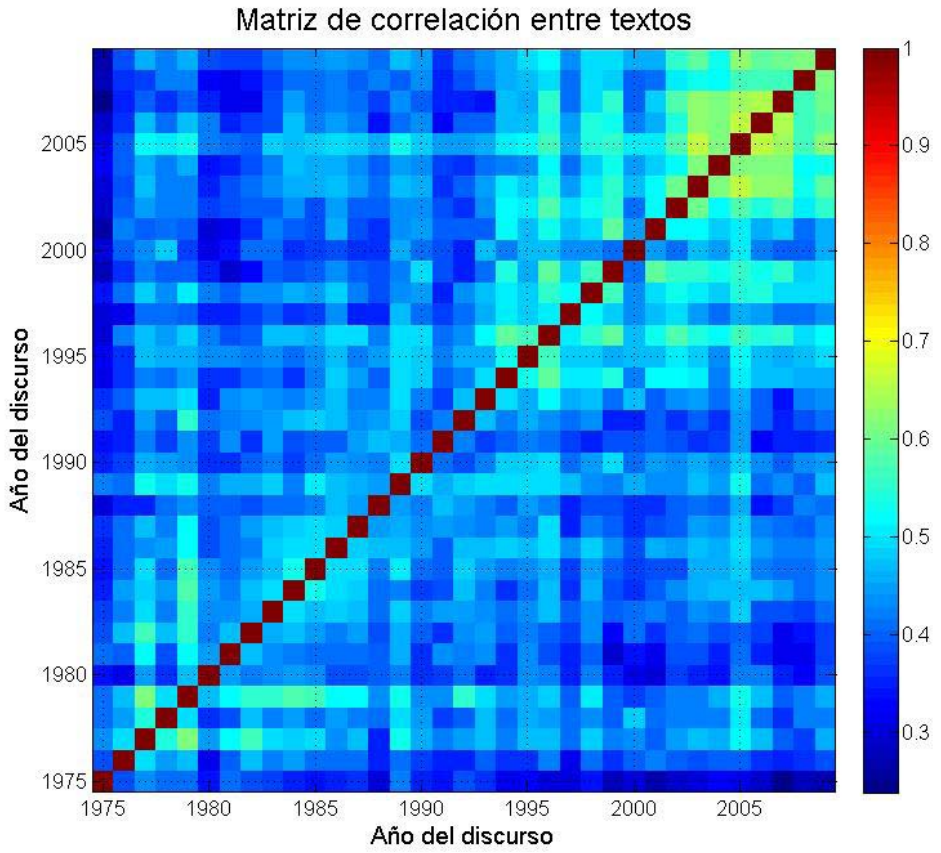
Análisis Semántico

- Latent Semantic Analysis
  - Estudio de la correlación entre textos atendiendo a la frecuencia de aparición de las palabras.
  - Preprocesado:
    - Filtrado de las palabras contenido (*stoplist*).
    - Utilización de funciones de ponderación y normalización.
  - Resultados:
    - Cuantificación y representación del contenido semántico.
    - Identificación de discursos anómalos.
    - Reconstrucción o filtrado de la semántica de los textos.

Análisis Semántico

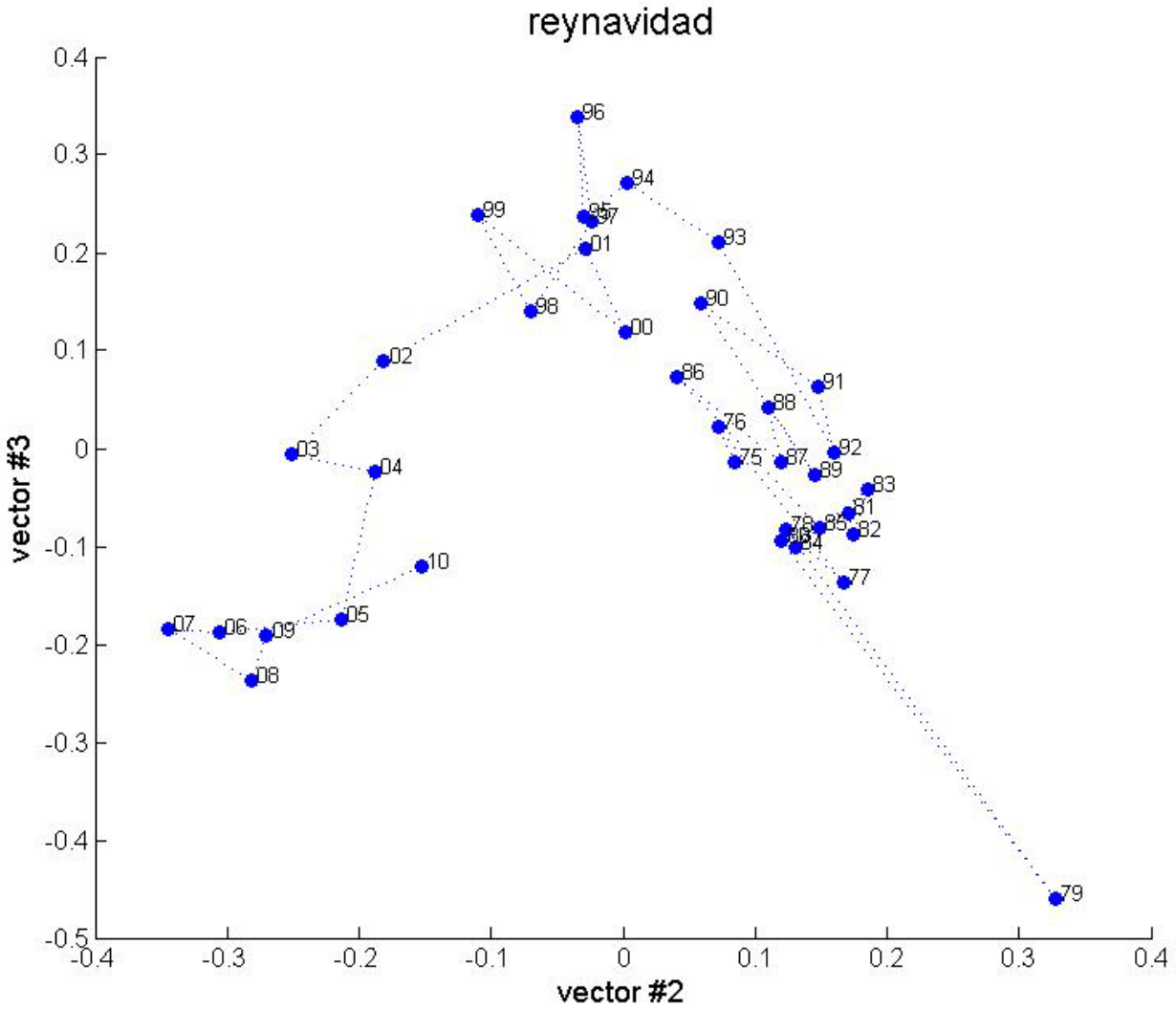
- Latent Semantic Analysis

Matriz de correlación de los discursos navideños del Rey Juan Carlos I.

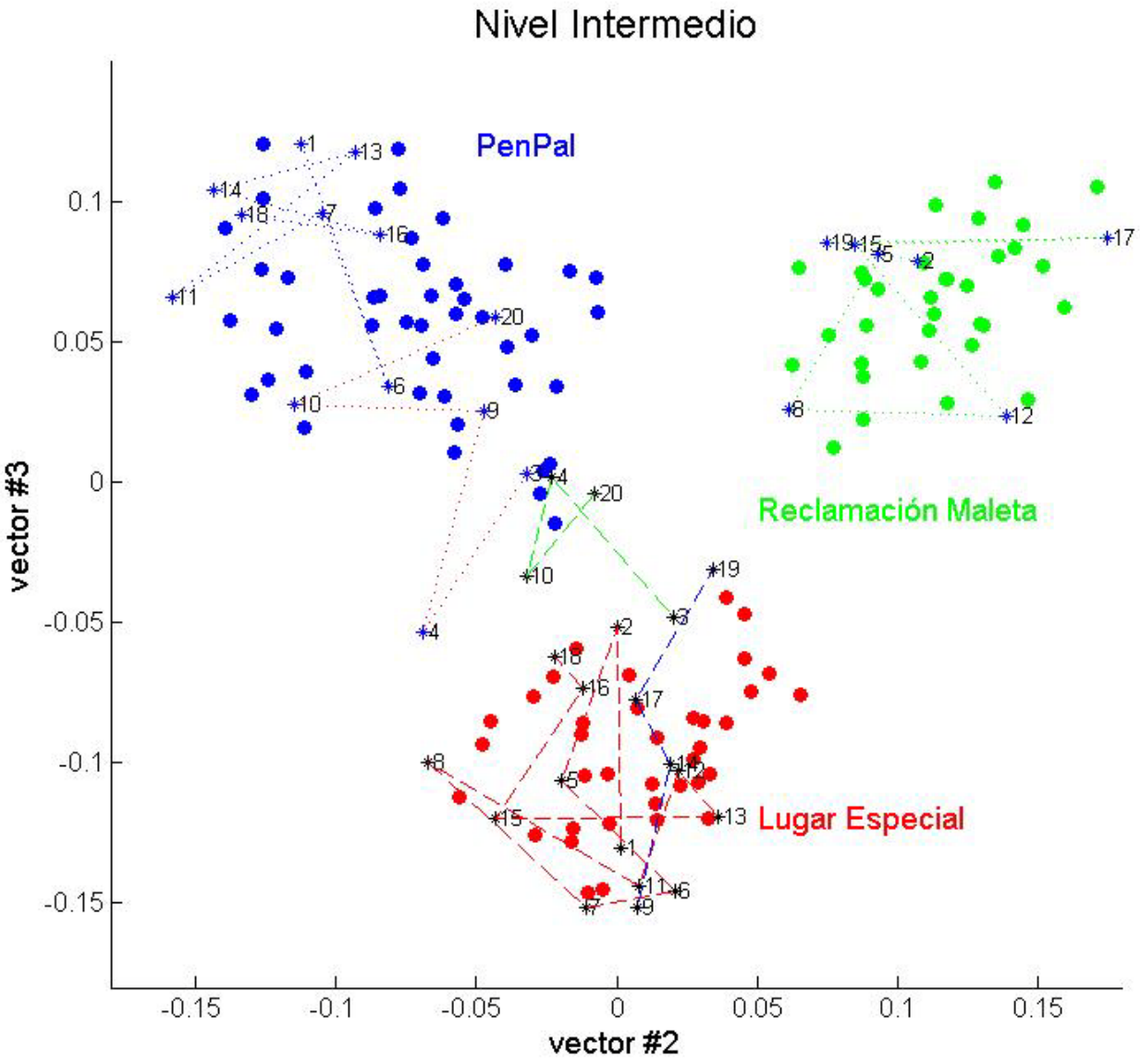


Análisis Semántico

- Latent Semantic Analysis: Discursos Navideños



LSA de los exámenes DELE Intermedio



Conclusiones

- Se han identificado textos automáticamente, sin leerlos.
- Se ha parametrizado el contenido semántico de un texto, comparativa e indivualmente, mediante la identificación de vocablos en ***campos semánticos*** definidos por el *PCIC*.
- Se han empleado herramientas de ***estadística multivariante,*** aplicándolas al ámbito de la lingüística y de la evaluación de una L2.
- Se obtiene valores para ***medir cuantitativamente*** el contenido esperado en un tipo de texto con una temática concreta.
- La identificación automática del contenido de un texto hoy es un paso adelante para la ***evaluación automática de textos***.

Identificación automática de la temática de un texto

¡Muchas gracias por su atención, ☺!

Paz Ferrero García de Jalón

paz.ferrero@uam.es

Universidad Autónoma de Madrid:

Departamento de Filología Inglesa

Universidad Popular de Alcobendas

21-23 de noviembre de 2011