

U LA Term Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Iria da Cunha Fanego
Instituto Universitario de Lingüística Aplicada (IULA)
Universidad Pompeu Fabra

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 1

U LA Term Integración de técnicas lingüísticas para el resumen automático de artículos médicos

GUIÓN

1. Introducción
2. ¿Qué es un resumen?
3. Estado de la cuestión en resumen automático
4. Nuestra aproximación
 - 4.1. Marco teórico de nuestra aproximación
 - 4.2. Corpus de análisis
 - 4.3. Nuestra idea central
 - 4.4. Hacia un resumen justificado lingüísticamente
 - 4.5. Validación de resultados
5. Trabajo futuro

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 2

U LA Term Integración de técnicas lingüísticas para el resumen automático de artículos médicos

1. INTRODUCCIÓN

SEMINARIO ULATERM 130606 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 3

U LA Term Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Motivación para la investigación en resumen automático

- Aumento de la información que nos llega en la actualidad. Ej. Internet.
- Necesidad de procesar esta información en el menor tiempo posible para tomar decisiones.
- Avanzado estado de la lingüística computacional (posibilidades reales de implementación).

SEMINARIO ULATERM 130606 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 4

U LA Term Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Nuestra aproximación

Intentar mejorar los actuales sistemas de resumen automático (basados en su mayoría en técnicas estadísticas) mediante la integración de técnicas lingüísticas, que no tomen los textos únicamente como entidades matemáticas.

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 5

U LA Term Integración de técnicas lingüísticas para el resumen automático de artículos médicos

2. ¿QUÉ ES UN RESUMEN?

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 6

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Resumen o *abstract*

- “Condensación de los conceptos principales del contenido del texto al que hace referencia” (Burgos *et al*, 1994).
- “An abbreviated, accurate representation of the contents of a document, preferably prepared by its author(s) for publication with it” (ANSI; Bathia, 1993).

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 7

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Parámetros de elaboración del resumen

- **Input:**
 - Único documento / Varios documentos
 - Dominio específico / Ámbito general
 - Texto monolingüe / Texto multilingüe
- **Output:**
 - Abstract / Extract
 - Resumen neutral / Resumen evaluativo
- Propósito del resumen:
 - Resumen indicativo / Resumen informativo
 - Necesidades del autor / Necesidades del usuario
 - Destinatario lego / Destinatario experto

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 8

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

3. ESTADO DE LA CUESTIÓN EN RESUMEN AUTOMÁTICO

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 9

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Métodos de resumen automático (Hovy, 2003)

- **Métodos superficiales:** basados en frecuencias de palabras, posición de determinados fragmentos, títulos, palabras clave,...
- **Métodos de nivel medio:** basados en el reconocimiento de cadenas léxicas, elementos relacionados, coreferencia, Máxima de Relevancia Marginal,...
- **Métodos profundos:** basados en la estructura del discurso.

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 10

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Método profundo: utilización de la estructura discursiva

- *Rhetorical Structure Theory* de Mann & Thompson (1988): estructura interna y relaciones discursivas de los textos.
- Marcu (1997): dependiendo del tipo de relación discursiva se seleccionarán o no determinados fragmentos para el resumen.

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 11

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

4. NUESTRA APROXIMACIÓN AL RESUMEN AUTOMÁTICO

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 12

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

4.1. Marco teórico de nuestra aproximación

13

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Marco teórico de nuestra aproximación

- 1) *Rhetorical Structure Theory*: Mann & Thompson (1988).
- 2) Sintaxis de Dependencias (*Meaning-Text Theory*): Mel'cuk (1988).

14

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

1) *Rhetorical Structure Theory (RST)*

- Mann & Thompson (1988); *Information Sciences Institute*, Universidad de California del Sur.
- Teoría descriptiva de organización del texto basada en su estructura interna y en sus relaciones discursivas (unidades discursivas mínimas: patrones núcleo-satélites).
- Representación mediante árboles jerárquicos de estructuras discursivas.

15

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Fragmento de estructura arbórea de relaciones discursivas

16

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Relación RST-Marcu

Marcu (1996-2000) relaciona la RST con la generación automática de resúmenes:

- segmentación del texto en unidades mínimas,
- relaciones de la RST,
- estructura retórica arbórea (núcleo-satélites),
- marcadores discursivos,
- algoritmo para *parsing* automático,
- aplicación al resumen automático.

17

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

2) Sintaxis de Dependencias

- Mel'cuk (1988); *Meaning-Text Theory*.
- Teoría caracterizada por concebir la sintaxis profunda como estructuras formadas por:
 - actantes (en español puede haber un máximo de seis: I-VI),
 - adjuntos (Attrib, Append, Coord).

18

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

“Ser niño se asoció con la utilización inadecuada del servicio de urgencias.”

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 19

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

“El estudio ha identificado como factores asociados al uso inadecuado de los servicios de urgencias la mejor accesibilidad geográfica al hospital, ser niño y acudir por iniciativa del paciente (frente a los enfermos enviados por su médico).”

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 20

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

4.2. Corpus de análisis

SEMINARIO ULATERM 130606 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 21

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Corpus de análisis

- 20 artículos médicos en español de la revista *Medicina Clínica*.
- Subcorpus del Corpus Técnico especializado del IULA – UPF: <http://brangaene.upf.es/bwananet/index.htm>

SEMINARIO ULATERM 130606 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 22

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Motivación de la elección del corpus

- Los artículos van acompañados de sus respectivos resúmenes que nos servirán posteriormente para comparar con los nuestros.
- Un ámbito concreto permitirá establecer generalizaciones más precisas: mismas estructuras, estilo y fenómenos discursivos e informativos.

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 23

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

4.3. Idea central de la tesis

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 24

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LATerm

Idea central de la tesis

Combinación de informaciones lingüísticas relevantes para llegar al resumen automático:

- estructura textual,
- estructura discursiva,
- estructura sintáctica.

SEMINARIO ULATERM 130605
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

25

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LATerm

Estrategia principal a seguir

- Combinar la estructura textual, discursiva (RST) y sintáctica (de dependencias) para lograr una representación más exacta del contenido del texto y solucionar problemas derivados de la simple utilización de la RST (Marcu, 1997) de cara al resumen.
- Crear una serie de reglas (a partir del análisis de nuestro corpus) que integren las perspectivas discursiva y sintáctica para que posteriormente puedan ser convertidas en un algoritmo (validación con un corpus de contraste).

SEMINARIO ULATERM 130605
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

26

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LATerm

Ejemplo de carencia del sistema de Marcu

[Smart cards have **two** main advantages over magnetic-stripe-card.³] [First, they can carry 10 or even 100 times as much information⁴] [-and hold it much more robustly.⁵] [Second they can execute complex tasks in conjunction with a terminal.⁶]

SEMINARIO ULATERM 130605
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

27

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LATerm

Ejemplo similar de nuestro corpus

[El análisis de regresión logística identificó **tres** variables asociadas, de forma independiente, con una visita apropiada a urgencias:¹] [acudir a este servicio por indicación de un médico²,] [vivir fuera de la región respecto a residir en la ciudad en la que está el hospital³] [y pertenecer a los grupos de consultas quirúrgicas y traumatismos respecto a la enfermedad médica y pediatría.⁴]

SEMINARIO ULATERM 130605
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

28

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LATerm

Nuestra solución

Elaboración

Núcleo: I-----I
Act I - Vb - Act II (con numeral)

Satélite: I-----I-----I-----I
ATTRIB

* Si tenemos esta estructura mantenemos el satélite

SEMINARIO ULATERM 130605
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

29

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LATerm

Árbol discursivo

15-18
elaboration-additional

(15) El análisis de regresión logística identificó tres variables asociadas, de forma independiente, con una visita apropiada a urgencias:

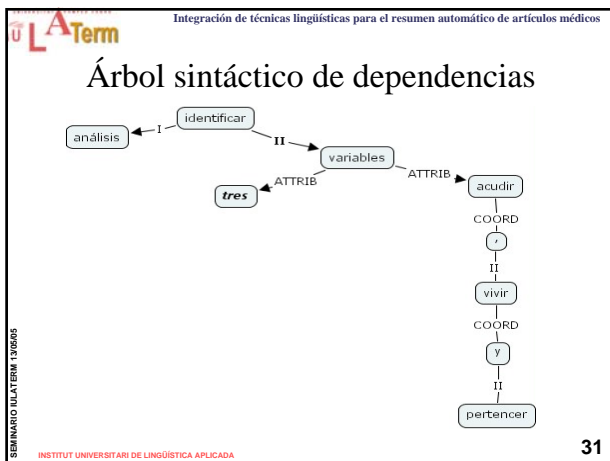
(16) acudir a este servicio por indicación de un médico.

(17) vivir fuera de la región respecto a residir en la ciudad en la que está el hospital

(18) y pertenecer a los grupos de consultas quirúrgicas y traumatismos respecto a la enfermedad médica y pediatría.

SEMINARIO ULATERM 130605
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

30



Integración de técnicas lingüísticas para el resumen automático de artículos médicos

4.4. Hacia un resumen justificado lingüísticamente

SEMINARIO ULATERM 130605
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

32

- Integración de técnicas lingüísticas para el resumen automático de artículos médicos
- ### a) Criterios para el resumidor
- 1- Criterios textuales.
 - 2- Criterios léxicos.
 - 3- Criterios sintáctico-discursivos.
 - 4- Coincidencia entre criterios léxicos y sintáctico-discursivos.
- SEMINARIO ULATERM 130606
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA
- 33

- Integración de técnicas lingüísticas para el resumen automático de artículos médicos
- ### 1. Criterios textuales
- Los artículos científicos y sus *abstracts* correspondientes pasan por cuatro *moves* para expresar el orden lógico del pensamiento científico (Salager-Meyer, 1991):
- Fundamento,
 - Pacientes y Métodos,
 - Resultados,
 - Discusión.
- SEMINARIO ULATERM 130606
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA
- 34

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

La estructura del artículo médico (I): Fundamento

- Es breve y debe proporcionar sólo la explicación necesaria para que el lector pueda comprender el texto que sigue a continuación.

SEMINARIO ULATERM 130605
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

35

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

La estructura del artículo médico (II): Pacientes y Métodos

- Indica el centro donde se ha realizado la investigación, el período de duración, las características de la serie estudiada, el criterio de selección empleado y las técnicas utilizadas.
- Proporciona los detalles suficientes para que una experiencia determinada pueda repetirse sobre la base de esta información.

SEMINARIO ULATERM 130605
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

36

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LA Term

La estructura del artículo médico (III): Resultados

- Se relatan (no interpretan) las observaciones efectuadas con el método empleado.

SEMINARIO ULATERM 130605

INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

37

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LA Term

La estructura del artículo médico (IV): Discusión

- Los autores tienen que exponer sus propias opiniones sobre el tema e interpretar los resultados.

SEMINARIO ULATERM 130605

INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

38

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LA Term

Estrategia de utilización de la estructura textual

- Seleccionar información de cada uno de los cuatro apartados para no perder información del proceso lógico de transmisión científica.

SEMINARIO ULATERM 130606

INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

39

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LA Term

Tabla de frecuencia de subtítulos en un corpus de 20 artículos médicos

Apart	Resumen	Frec	Artículo	Frec
1	Fundamento Objetivo	18 2	----	20
2	Pacientes y métodos Sujetos y métodos Material y método Métodos Población y métodos	8 4 2 3 3	Pacientes y métodos Sujetos y métodos Material y método Método Poblaciones y método	11 3 2 1 3
3	Resultados	20	Resultados	20
4	Conclusiones	20	Discusión	20

SEMINARIO ULATERM 130606

INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

40

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LA Term

2. Criterios léxicos

- Identificación de unidades léxicas (básicamente sustantivos y verbos) indicadores de relevancia en el contexto de la investigación médica.
- Búsquedas de contextos con Bwananet.

SEMINARIO ULATERM 130605

INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

41

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

LA Term


Unidades léxicas

- Sustantivos: *objetivo, objeto, propósito, intención, resumen, conclusión, resultado, estudio, trabajo.*
- Verbos: *realizar, asociar, analizar, presentar, evaluar, relacionar, aportar, estudiar, valorar, incluir, observar, llevar a cabo, obtener, alcanzar, encontrar.*

SEMINARIO ULATERM 130605

INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

42



 Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Estrategia de utilización de unidades léxicas

- Seleccionar las oraciones que contengan estas unidades léxicas.
- Posteriormente se compararán estas oraciones con las seleccionadas mediante criterios sintáctico-discursivos para comprobar su coincidencia.

43

SEMINARIO ULATERM 130605
 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA



 Integración de técnicas lingüísticas para el resumen automático de artículos médicos

3. Criterios sintáctico-discursivos

- Aplicación de las reglas sintáctico-discursivas.
- Selección de los fragmentos de texto relevantes.

44

SEMINARIO ULATERM 130605
 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA



 Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Ejemplo de regla (I)

- Si hay un satélite de Elaboración, que sea un atributo (ATTRIB), y que además sea una oración de relativo explicativa, se elimina dicho satélite.

45

SEMINARIO ULATERM 130606
 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA



 Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Ejemplo

“Coincidiendo con ese mismo estudio, la visita previa a un médico es un factor asociado a una mayor adecuación del uso del servicio de urgencias, [lo que estaría en relación con el papel de filtro de la atención primaria.]”

46

SEMINARIO ULATERM 130606
 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA



 Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Ejemplo de regla (II)

- Si hay un satélite de Elaboración que además sea un APPEND, se elimina dicho satélite.

47

SEMINARIO ULATERM 130605
 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA


 Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Ejemplo

“[Coincidiendo con ese mismo estudio], la visita previa a un médico es un factor asociado a una mayor adecuación del uso del servicio de urgencias, lo que estaría en relación con el papel de filtro de la atención primaria.”


48

SEMINARIO ULATERM 130605
 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Reducción

“Coincidiendo con ese mismo estudio, la visita previa a un médico es un factor asociado a una mayor adecuación del uso del servicio de urgencias, lo que estaría en relación con el papel de filtro de la atención primaria.”



“La visita previa a un médico es un factor asociado a una mayor adecuación del uso del servicio de urgencias.”

49

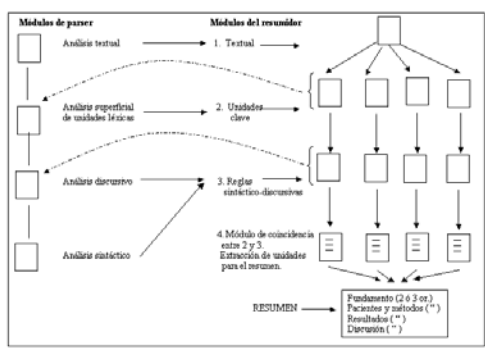
Integración de técnicas lingüísticas para el resumen automático de artículos médicos

4. Coincidencia entre criterios léxicos y sintáctico-discursivos

- Si coinciden las oraciones resultantes de la aplicación de los criterios léxicos y sintáctico-discursivos: éstas son seleccionadas para el resumen.
- Si no coinciden las oraciones: toma de decisiones.

50

Integración de técnicas lingüísticas para el resumen automático de artículos médicos



51

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

b) Módulos del analizador

- *Parser* morfosintáctico:
 - tokenizador,
 - tagger,
 - lematizador,
 - analizador sintáctico de dependencias (limitación: no existe en español. En curso: Montse Marimón. Posibilidad de adaptación).
- Analizador léxico: desambiguador.
- *Parser* discursivo (limitación: no existe en español. En inglés: Marcu).

52

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

4.5. Validación de resultados

53

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Resumen del autor como ideal

- Especialista en la materia.
- Mismo autor del artículo y del resumen.
- Resumen orientado a especialistas.
- Guía de estilo para publicaciones médicas.
- Cuatro apartados fijados (IMRD).
- Publicación del artículo CON el resumen.

54

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Justificación estadística

- Prueba con 3 médicos y 3 lingüistas para corroborar que el resumen del autor es correcto.
- Corpus de 20 artículos de la revista *Medicina Clínica* (sin el resumen).
- Comparación de los contenidos del resumen del autor con los de los resúmenes de 3 médicos y de 3 lingüistas: ANÁLISIS *CLUSTER*

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 55

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Instrucciones a 3 médicos y 3 lingüistas

- **Subraye** en cada texto las oraciones que considere *indispensables* para construir un buen resumen del mismo (excepto títulos y subtítulos):
 - Originales = máximo 20 líneas subrayadas
 - Originales Breves = máximo 15 líneas subrayadas
- **Escriba** a continuación un resumen del texto:
 - Originales = máximo 250 palabras (aprox. 20 líneas)
 - Originales Breves = máximo 180 palabras (aprox. 15 líneas)

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 56

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

5. TRABAJO FUTURO

SEMINARIO ULATERM 130606 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 57

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Tareas futuras

- Ampliación del corpus.
- Desarrollo de más reglas que cubran todos los casos posibles.
- Métodos de evaluación.

SEMINARIO ULATERM 130606 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 58

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Evaluación prevista

- Comparación de nuestro resumen con los de sistemas comerciales: ANOVA.
- Comparación de nuestro resumen con el del autor: ANÁLISIS *CLUSTER*
- Comparación de nuestro resumen con los de los 3 médicos: ANÁLISIS *CLUSTER*

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 59

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Integración de técnicas lingüísticas para el resumen automático de artículos médicos

Iria da Cunha Fanego
 Instituto Universitario de Lingüística Aplicada (IULA)
 Universidad Pompeu Fabra

SEMINARIO ULATERM 130605 INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA 60