UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

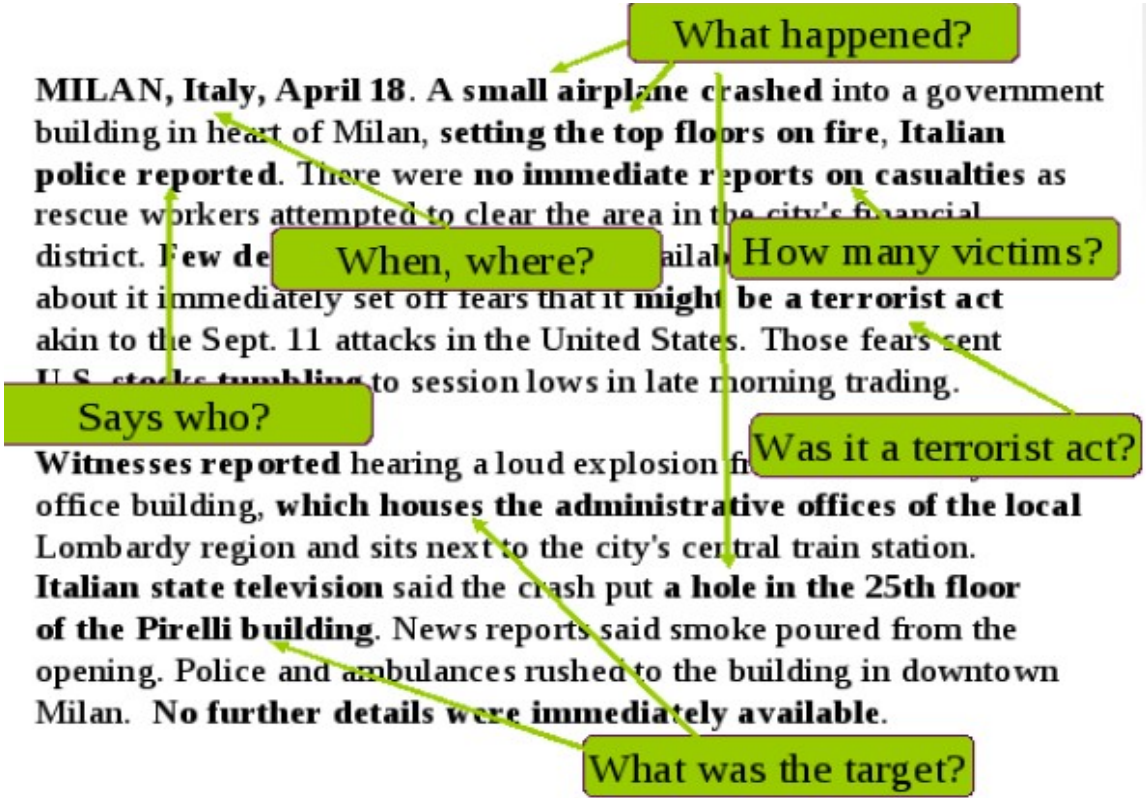
Seminario de análisis del discurso

Dr. César Antonio Aguilar
Facultad de Lenguas y Letras
08/11/2010

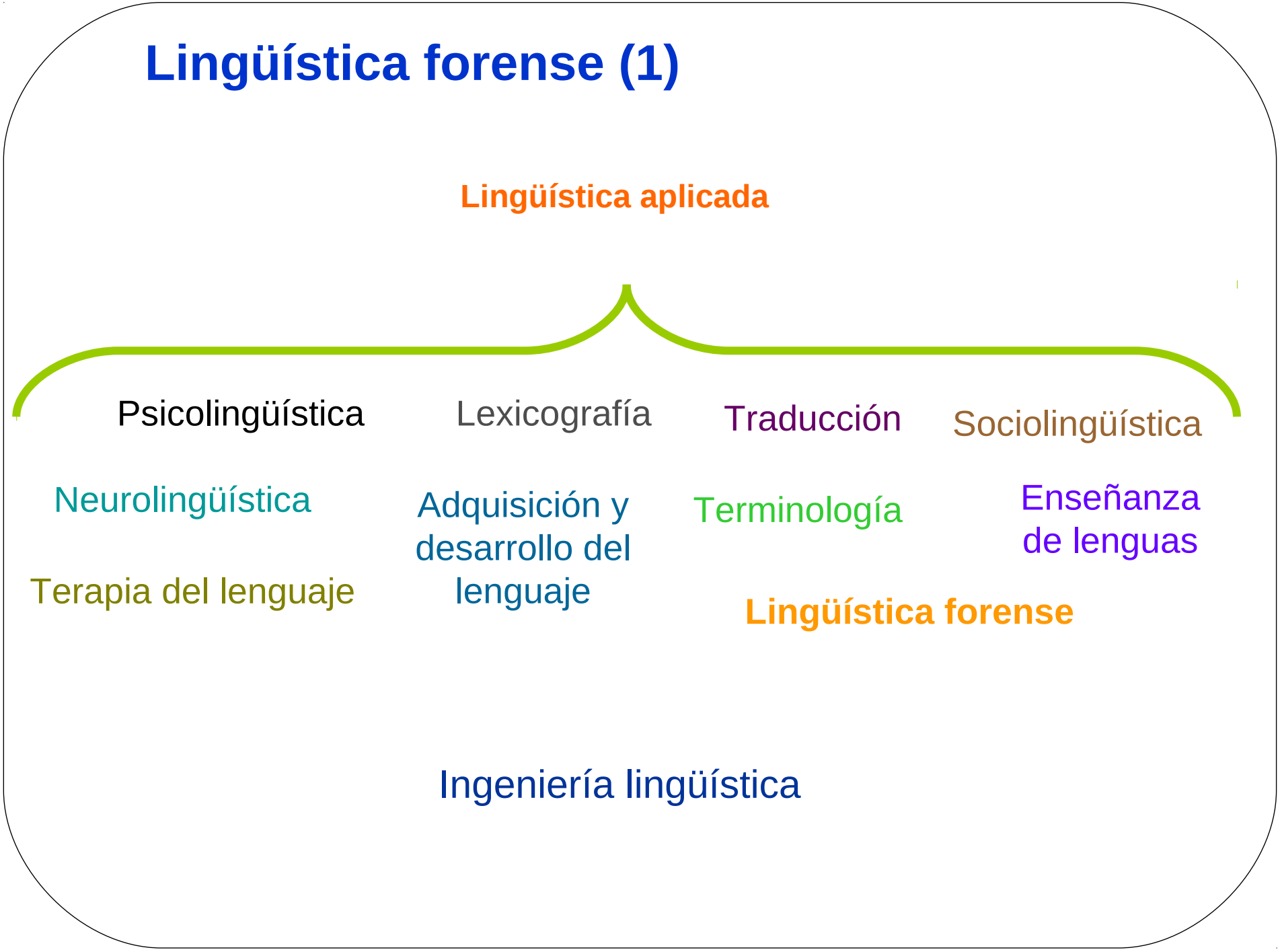
CAguilar@iingen.unam.mx

Aplicaciones prácticas del análisis del discurso

En esta clase, vamos a ver un ejemplo de aplicación de análisis del discurso para resolver un problema concreto: la detección de fragmentos plagiados entre textos, considerando algunas cuestiones propias de la lingüística aplicada y la ingeniería lingüística.



Lingüística forense (1)



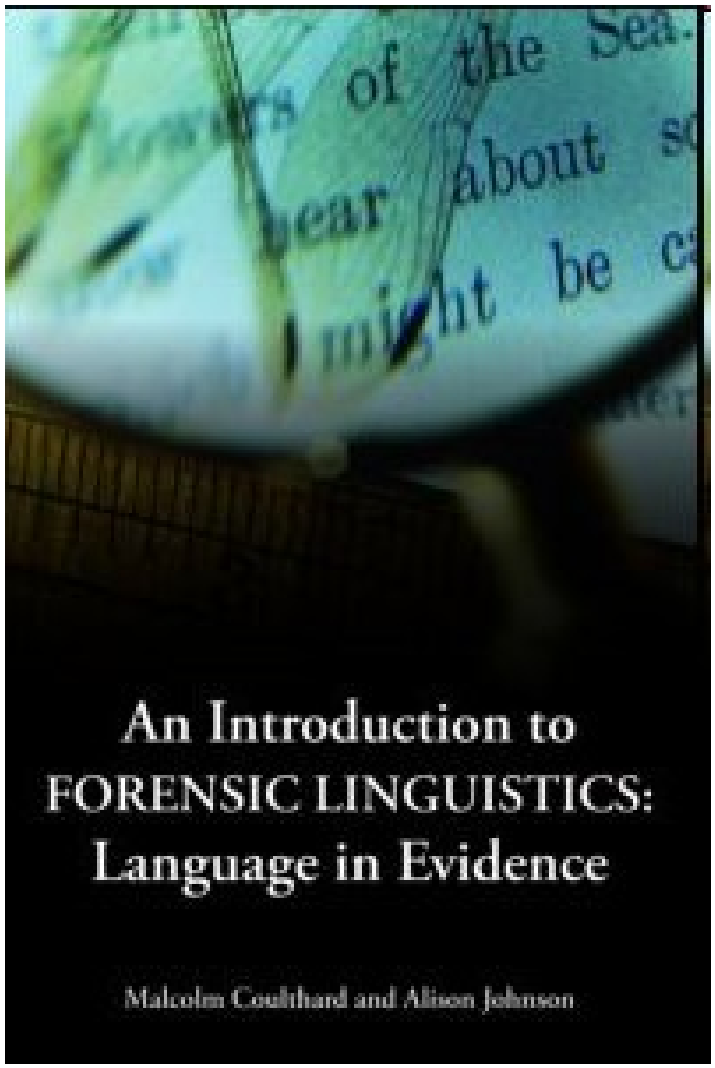
Lingüística forense (2)

Rama de la lingüística aplicada

Empleo de modelos y métodos de análisis lingüísticos para la obtención y valoración de evidencias en problemas legales

- Delimitación del sentido de una ley
- Evidencia lingüística
- Plagio
- Juicios orales

Lingüística forense (3)



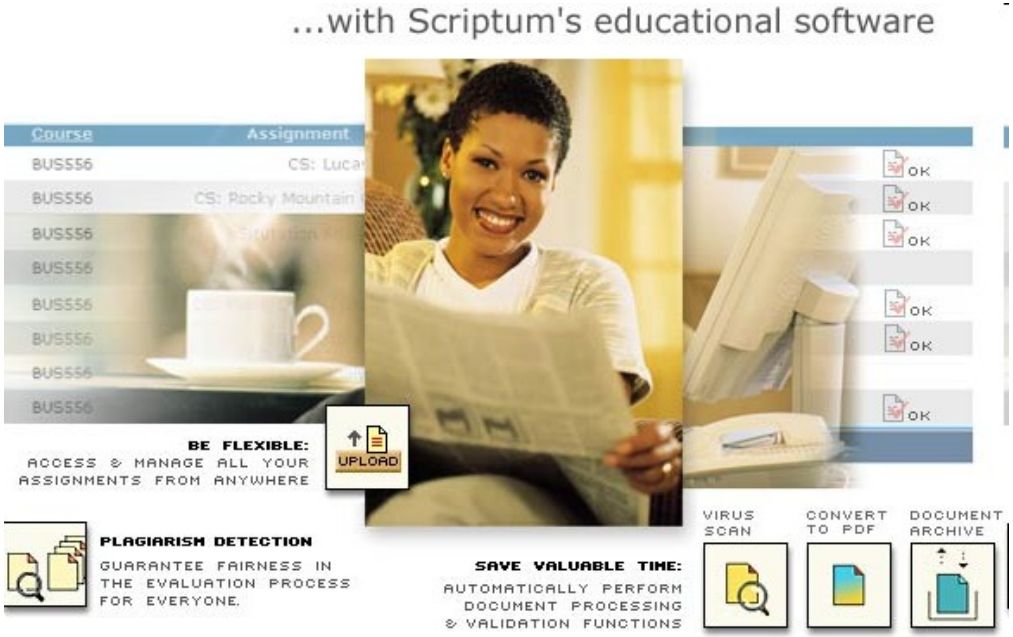
Evidencia lingüística para resolver problemas legales:

- Conversaciones
- Mensajes
- Correos electrónicos
- Grafología

Lingüística forense (4)

Detección automática de plagio:

- Artículos científicos
- Trabajos escolares
- Tesis
- Programas de cómputo
- Paráfrasis

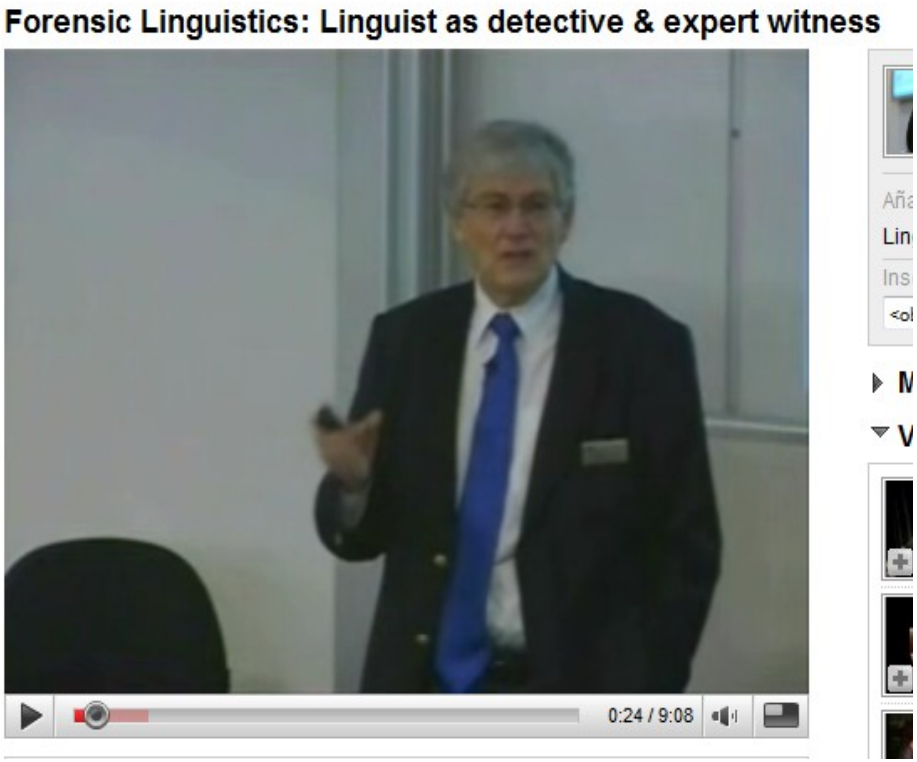


Método de análisis (4)

- La BBC News dedicó en 2004 un reportaje a la lingüística forense, titulado *Reading between the lines.*

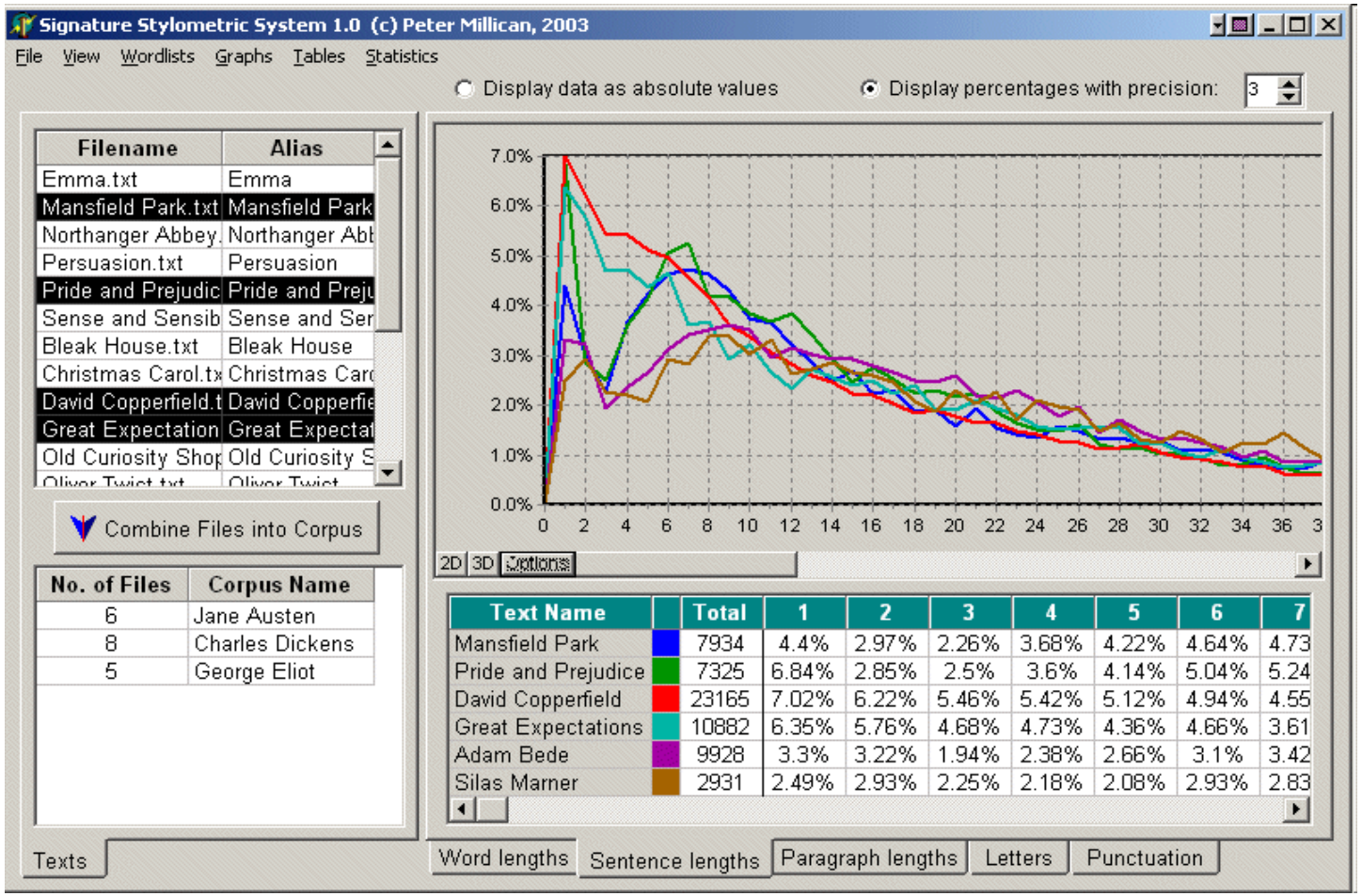
- Una buena explicación sobre la utilidad de la lingüística forense la da el Dr. Malcom Coulthard.

El video lo pueden ver en:



http://es.youtube.com/watch?v=4z6Krsjwc84

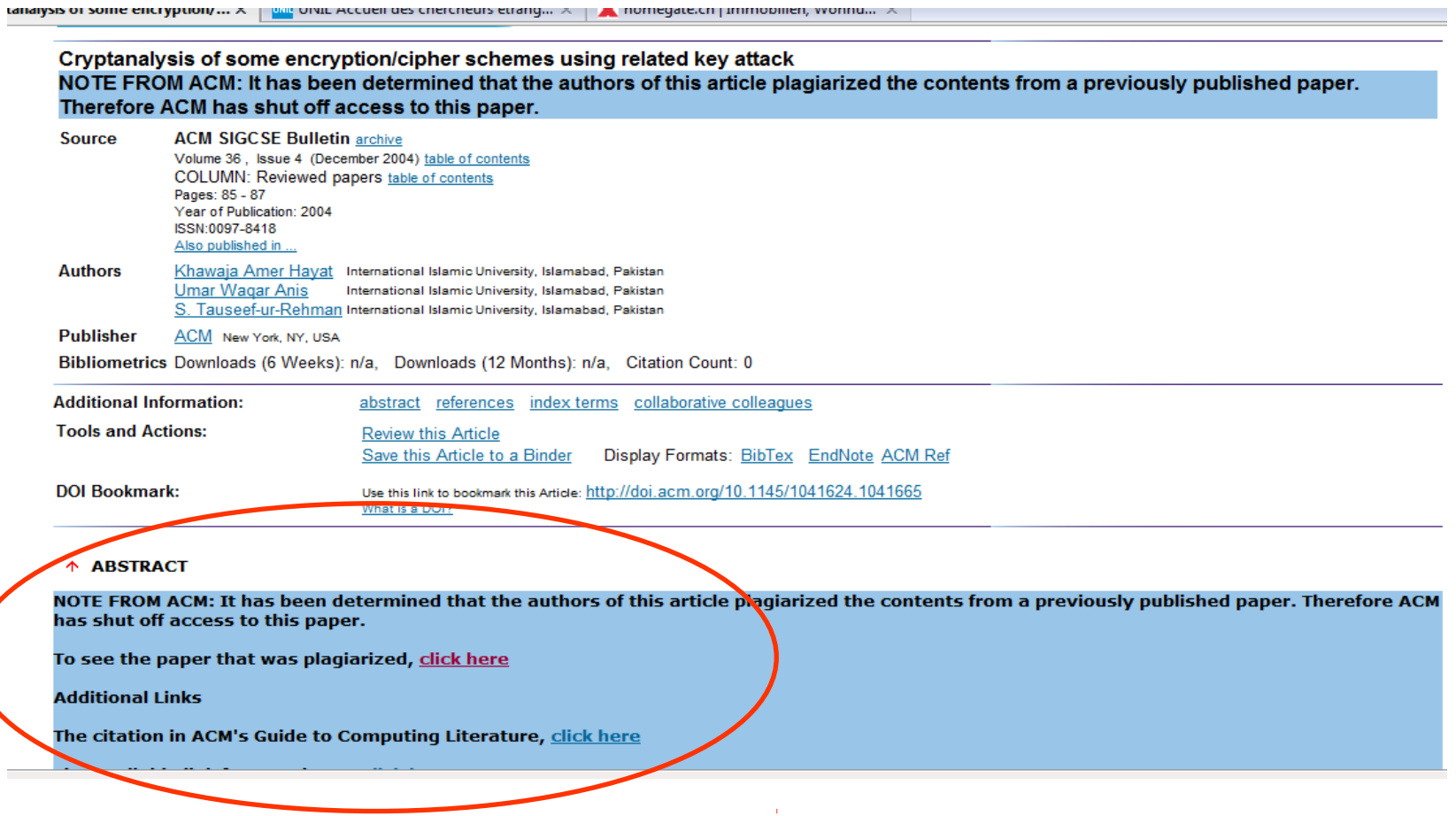
Método de análisis (5)



Existen herramientas que apoyan las tareas de búsqueda de evidencia lingüística que echan mano de métodos computacionales, estadísticos, lingüísticos, etc.

Método de análisis (6)

El impacto de la lingüística forense es una realidad que puede verse en ámbitos académicos, empresariales y gubernamentales.



Cryptanalysis of some encryption/cipher schemes using related key attack

NOTE FROM ACM: It has been determined that the authors of this article plagiarized the contents from a previously published paper. Therefore ACM has shut off access to this paper.

Source: ACM SIGCSE Bulletin archive
Volume 36 , Issue 4 (December 2004) table of contents
COLUMN: Reviewed papers table of contents
Pages: 85 - 87
Year of Publication: 2004
ISSN:0097-8418
Also published in ...

Authors: Khawaja Amer Hayat International Islamic University, Islamabad, Pakistan
Umar Waqar Anis International Islamic University, Islamabad, Pakistan
S. Tauseef-ur-Rehman International Islamic University, Islamabad, Pakistan

Publisher: ACM New York, NY, USA

Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Citation Count: 0

Additional Information: abstract references index terms collaborative colleagues

Tools and Actions: Review this Article
Save this Article to a Binder Display Formats: BibTex EndNote ACM Ref

DOI Bookmark: Use this link to bookmark this Article: http://doi.acm.org/10.1145/1041624.1041665
What is a DOI?

ABSTRACT

NOTE FROM ACM: It has been determined that the authors of this article plagiarized the contents from a previously published paper. Therefore ACM has shut off access to this paper.

To see the paper that was plagiarized, click here

Additional Links

The citation in ACM's Guide to Computing Literature, click here

Detección de fraude (1)

Hay información que se transmite en el trabajo, y cuenta con una estructuración lingüística específica.

¿Qué pasa si se filtra información “sospechosa”?

Detección de fraude (2)

Mensaje con información esperada:

¿Cuál es tu perspectiva sobre el financiamiento en moneda extranjera?
¿Cómo afecta el costo del swap ?

Mensaje con información “personal” (alude a un estado de ánimo):

Voy a pedir al fulanito que se transparenten claramente todos los problemas identificados, que se defina y tome decisiones al respecto, y que le dé seguimiento a los acuerdos que se vayan tomando y resultados que se vayan generando al día. Las soluciones no son tan simplistas. Estoy preocupado.

Detección de fraude (3)

Mensaje con información sospechosa :

Ahora bien, con respecto al otro tema de hacer un intercambio entre la bolsa para la basura vieja (BB1) y armar una nueva (BB2), aunque suena bien, creo que no debemos hacerlo, pues después de pensarlo bien, creo que tiene algunos riesgos fuertes. La BB1, aunque nos da problemas, ahorita está bien escondida.

Metodología

- Detección y análisis de palabras y patrones textuales claves.
- Búsqueda de dichas palabras y patrones en un corpus de correos electrónicos.
- Cálculo de frecuencias respecto a la ocurrencia de tales palabras y patrones.
- Clasificación de correos con base en las palabras y patrones detectados.

Análisis (1)

Palabras claves

- RESERVAR
- BASURA
- PUMPKIN
- PUFF
- EFECTO
- CÁLCULO
- DISMINUCIÓN
- AUMENTO
- GENERAR
- IMPACTO
- PORTAFOLIO
- MONTO
- BUCKET
- AJUSTAR
- RESERVAR
- CARTERA
- COBRANZA
- ADQUISICIÓN
- SUSTITUCIÓN
- VENTA
- MILLÓN
- BOLSA

Frecuencia

SUSTITU:	40.55685094218655072647
AUMENT:	40.45240102710377934621
ADQUISICI:	36.90380116581989959696
INCREMENT:	31.74894503469579314616
PUFF:	30.751425603744844486 50
VENTA DE CARTERA:	30.22867811660613996828
BUCKET:	29.717594438863510807 02
PORTAFOLIO:	28.39429726266523687026
EFECTO:	27.611592253190467952 88
COMPRA:	26.298872112178914200 10
AJUST:	23.596853814144075721 84
GENERAR:	22.216580759578287430 74
IMPACTO:	21.812387568612831162 96
RESERV:	21.064153242099656144 88
MONTO:	20.743003111309846636 41
COBRANZA:	20.311828223616672887 34
CARTERA:	20.308135915872577402 29
PUMPKIN:	19.049919479740847645 68
DISMINU:	17.020098664726198133 73
BASURA:	15.362756769355767698 97

Análisis (2)

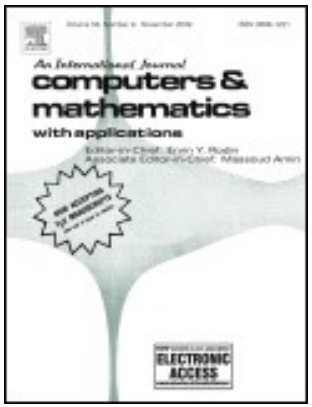
Correos electrónicos enviados durante 30 días

NOMBRE	DE	PARA	CC	TOTAL
Sujeto 1	0	0	6	6
Sujeto 2	0	7	0	7
Sospechoso 1	30	42	12	84
Sujeto 3	0	3	0	3
Sujeto 4	0	2	13	15
Direccion	0	0	7	7
Sujeto 5	0	0	1	1
Sujeto 6	1	6	0	7
Sujeto 7	4	8	0	12
Sujeto 8	1	7	0	8
Sospechoso 2	16	26	12	54
Sujeto 9	0	0	1	1
Sujeto 10	16	31	13	60

Plagio y paráfrasis (1)

Dos autores abordan un problema y presentan dos soluciones diferentes. La duda es que el texto de uno (Autor Y), se parece mucho al del otro (Autor X).

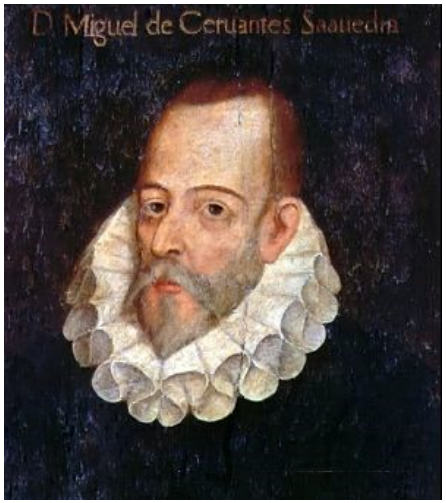
Plagio y paráfrasis (2)

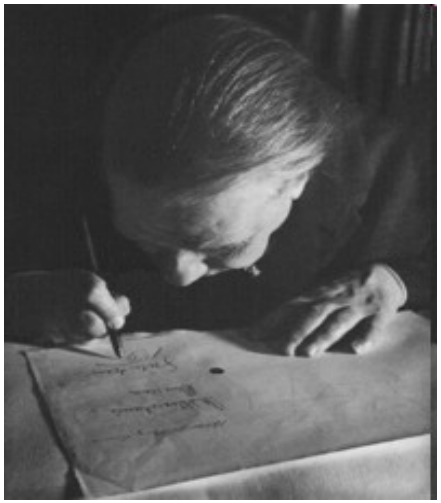


¿Qué es exactamente “plagiar”?

Si elaboro un artículo en donde ocupo la noción de entropía aplicada a una tarea de extracción de información, y no doy mayores referencias (justo porque explicarlo sería equivalente a decir algo completamente trivial), ¿estoy cometiendo plagio?

Incluso, hay muchos autores que se “autoplagian” párrafos de sus textos, o por lo menos los parafrasean para crear nuevos documentos. ¿Es malo hacer esto?





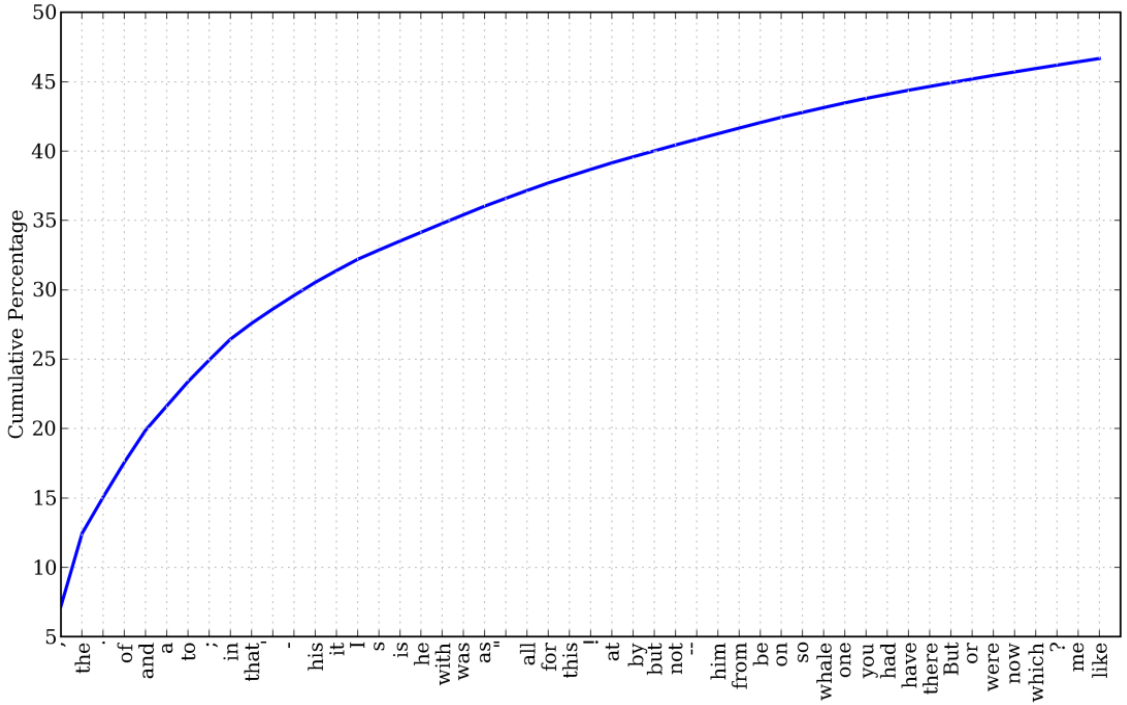
Plagio y paráfrasis (3)

Gustav Herdan (1897-1926) señala que el lenguaje cuenta con rasgos inherentes al estilo de una persona. *Ergo:*

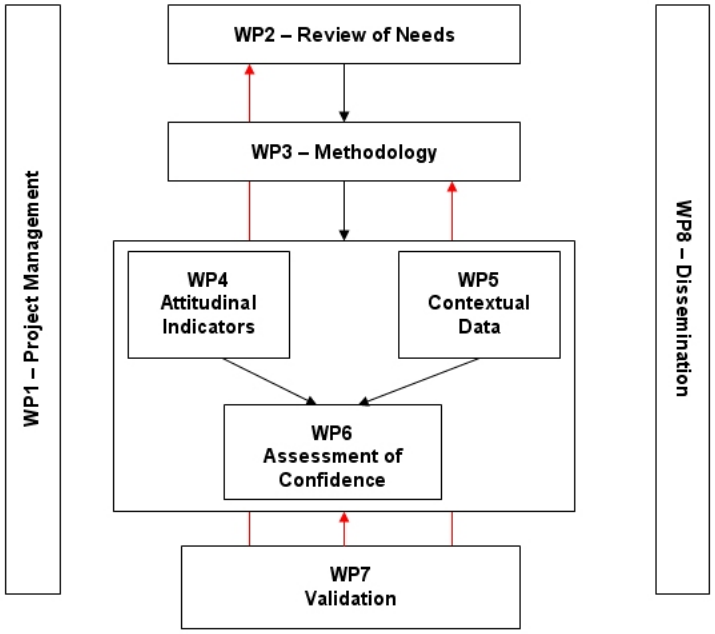
- Es algo que puede rastrearse matemáticamente.

- Establece contrastes entre palabras funcionales y de contenido.

- Lo más importante: nuestro vocabulario de palabras funcionales configura una “huella lingüística” nítida de nuestra persona.



Plagio y paráfrasis (4)



Si queremos rastrear por qué los autores X y Y se parecen, ¿podríamos comparar su frecuencia de uso de palabras de contenido y funcionales?

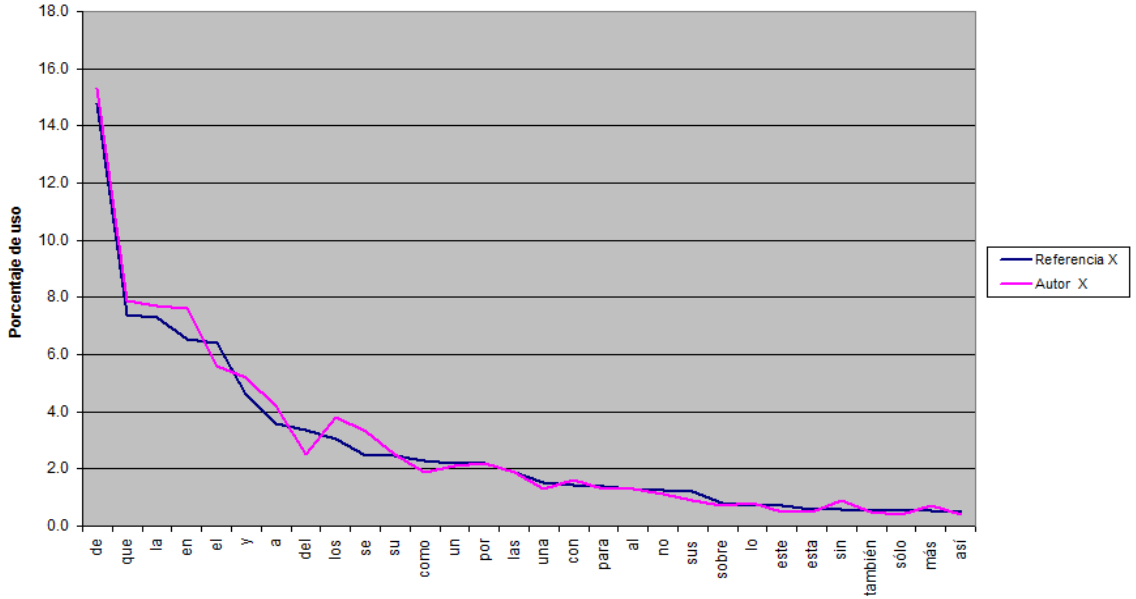
Juguemos a ser detectives: los autores X y Y, de entrada, pueden tener estilos propios, que no necesariamente deben parecerse.

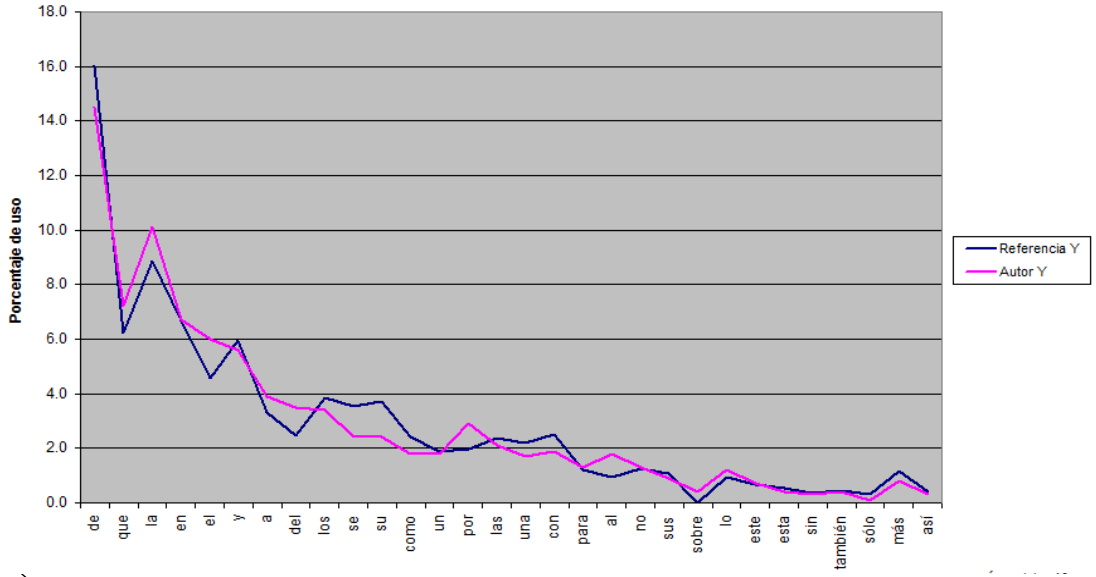
Si es así, ¿por qué justo el texto del autor Y se parece al de X?

Un problema más: supongamos que X publicó su texto antes que Y, y Y lo sabe. ¿El parecido es mera coincidencia, o no?

Plagio y paráfrasis (5)

Análisis de palabras funcionales de ***X*** con un texto anterior

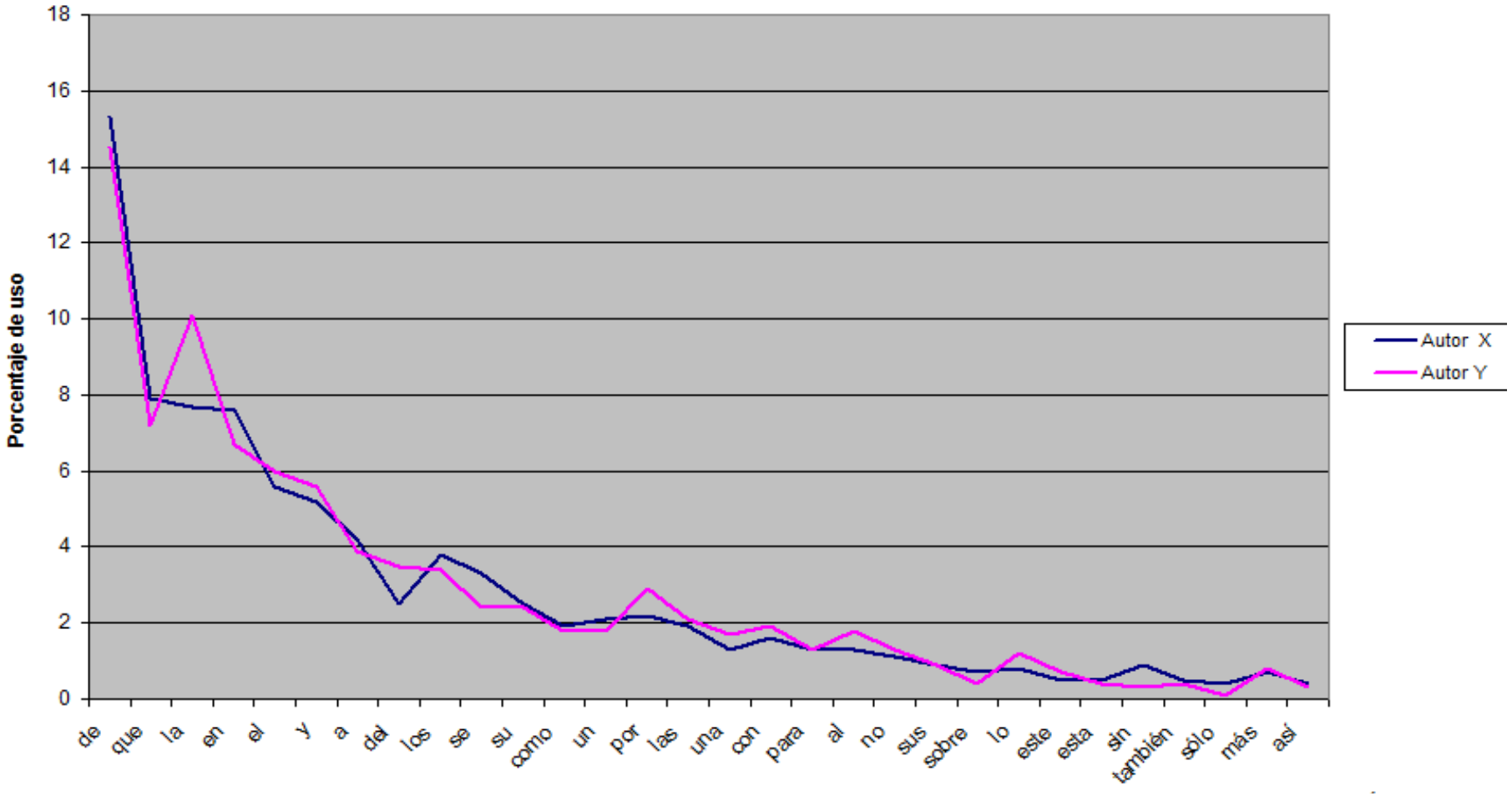




Análisis de palabras funcionales de ***Y*** con un texto anterior

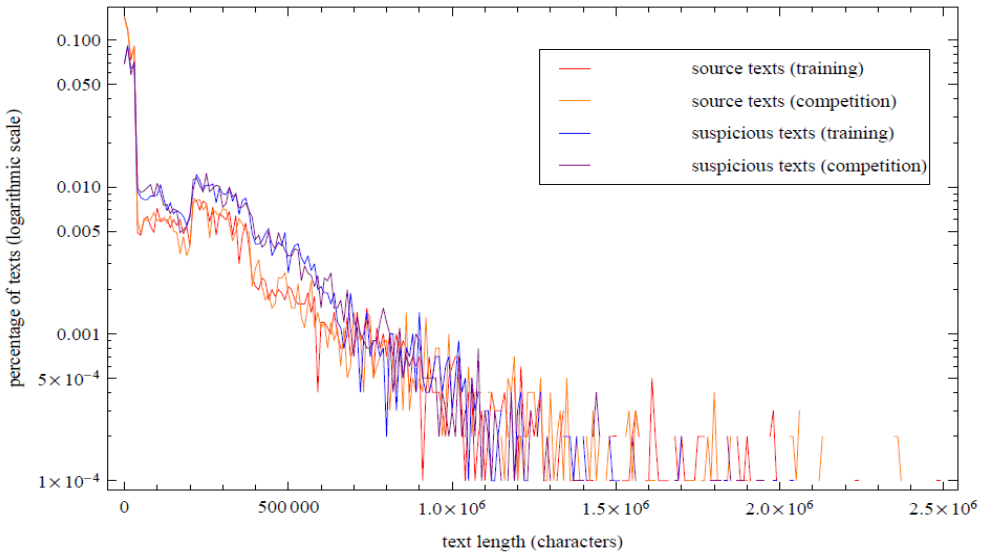
Plagio y paráfrasis (6)

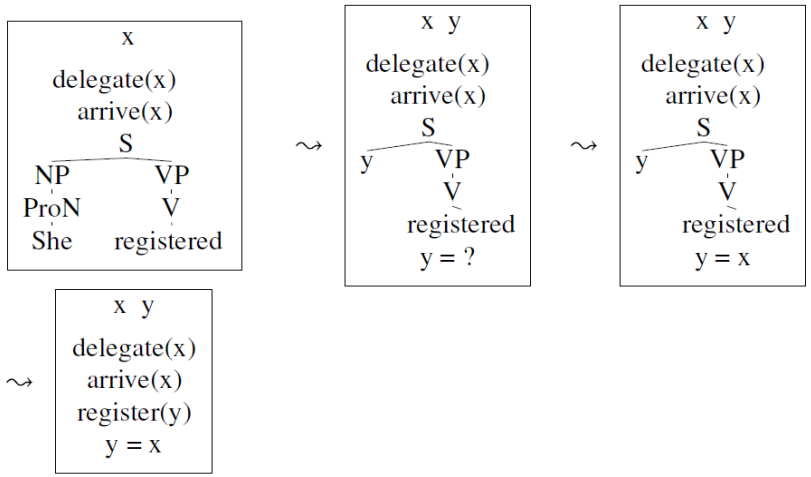
Análisis de palabras funcionales de *X* y *Y* entre sí



Comentarios finales

Corpus statistics





Problemas:

1. Tenemos buenos métodos y sistemas computacionales para procesar textos, pero no tenemos todavía buenos criterios para decidir qué es un plagio.

1. Así, parece que esta tarea, y otras propias de la lingüística forense, lo único que aportan realmente son evidencias lingüísticas para resolver casos.

1. ¿Esto es una tarea de analistas del discurso: sí o no, y por qué?

Gracias por su atención

Blog del curso: http://discurso-uaq.weebly.com/