

# Towards the 2nd Intl. Competition on Plagiarism Detection and Beyond

Alberto Barrón-Cedeño and Paolo Rosso

Natural Language Engineering Lab. - ELiRF  
Department of Information Systems and Computation  
Universidad Politécnica de Valencia, Spain  
{lbarron, proso}@dsic.upv.es

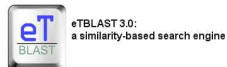
Plagiarism Conference  
June, 2010



Language Engineering Research Group  
**NLEL**  
Natural Language Engineering Lab



# The Plagiarism Resource Site



The Sherlock Plagiarism Detector

SID - Plagiarism Detection





**Which plagiarism detection method performs best?**



Language and Learning  
**NLEL**  
 National League for Engineering and Technology



# Outline

Introduction

Plagiarism Detection Overview

Preparing a Competition on Plagiarism Detection

First International Competition on Plagiarism Detection

Second International Competition on Plagiarism Detection

And beyond



Language Learning Technology  
**NLEL**  
Support Language Engineering Life

# Introduction: Automatic Plagiarism Detection

**Given a suspicious document  $d_q$ ...**

- detecting potential cases of plagiarism;
- if possible, providing the alleged source.



Language and Learning  
Research  
**NLEL**  
Support Learning. Empowering Life.

# Introduction: Automatic Plagiarism Detection

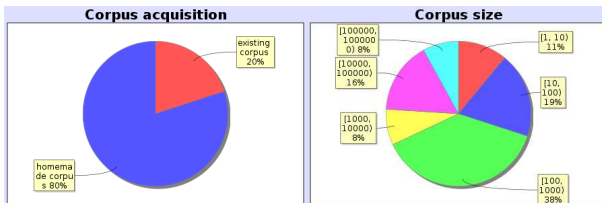
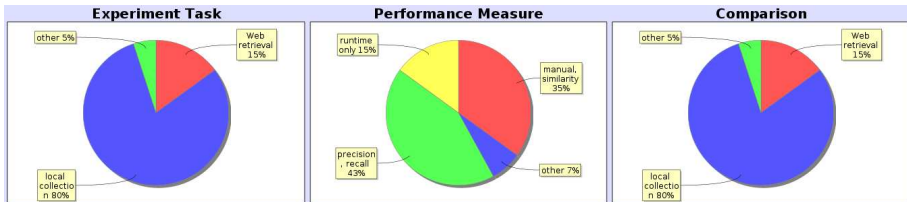
**Given a suspicious document  $d_q$ ...**

- detecting potential cases of plagiarism;
- if possible, providing the alleged source.

Afterwards, a person can take the final decision: whether a text is plagiarised or not.



# Introduction: Previous evaluations



[Potthast, et al., 2010, to appear]

# Outline

Introduction

Plagiarism Detection Overview

Preparing a Competition on Plagiarism Detection

First International Competition on Plagiarism Detection

Second International Competition on Plagiarism Detection

And beyond



Language Learning Technology  
**NLEL**  
Support Learning Engineering Life

# Plagiarism Detection Overview

Goal Identifying the plagiarized sections in a suspicious document  $d_q$ .



Language Learning through  
Research and Innovation  
**NLEL**  
Support Learning. Empowering Life.

# Plagiarism Detection Overview

Goal Identifying the plagiarized sections in a suspicious document  $d_q$ .

Objective Providing experts with evidence to decide whether a case of plagiarism is at hand.



# Plagiarism Detection Overview

Goal Identifying the plagiarized sections in a suspicious document  $d_q$ .

Objective Providing experts with evidence to decide whether a case of plagiarism is at hand.

Approaches

- *intrinsic*
- *external*



# PDO: Intrinsic Plagiarism Detection



An expert is often able to detect plagiarism by reading a document

Insertion of text from a different author into  $d_q$  causes **style** and **complexity** irregularities

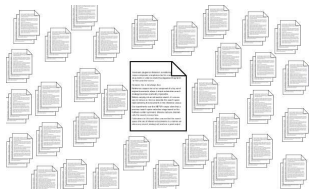
[Meyer zu Eißsen and Stein, 2006], [Stamatatos, 2009]



Language Center  
NLEL  
Support Learning Engineering Life



# PDO: External Plagiarism Detection



Better evidence than style irregularities is if the source of plagiarism case can be provided

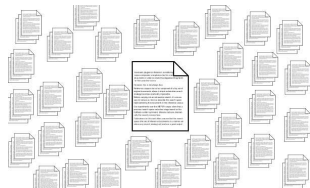
It is closer to information retrieval

[Potthast et al., 2009]



Language Engineering  
Research Center  
**NLEL**  
Support Language Engineering Life

# PDO: External Plagiarism Detection



Better evidence than style irregularities is if the source of plagiarism case can be provided

It is closer to information retrieval

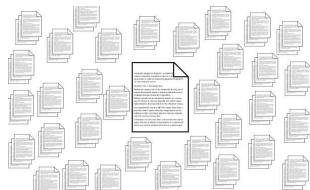
$d_q$  and a collection of potential source documents  $D$  are given. The task is to identify the plagiarized sections in  $d_q$  (if there are any), and their respective source sections in  $D$

[Potthast et al., 2009]



Language and Learning Technology  
**NLEL**  
Support Learning, Empowering Life

# PDO: External Plagiarism Detection



Better evidence than style irregularities is if the source of plagiarism case can be provided

It is closer to information retrieval

## Issues render this task difficult

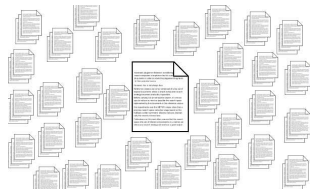
- Number of potential source documents,  $|D|$ ;
- Plagiarizing a text often includes paraphrasing, summarizing, and even translation.

[Potthast et al., 2009]



Language and Learning Technology  
NLEL  
Support Learning Engineering Life

# PDO: External Plagiarism Detection



Better evidence than style irregularities is if the source of plagiarism case can be provided

It is closer to information retrieval

## Models

Vector Space Models

Fingerprinting techniques

[Broder, 1997], [Maurer et al., 2006]

SPEX [Bernstein and Zobel, 2004],

Winnowing [Schleimer et al., 2003]

[Potthast et al., 2009]



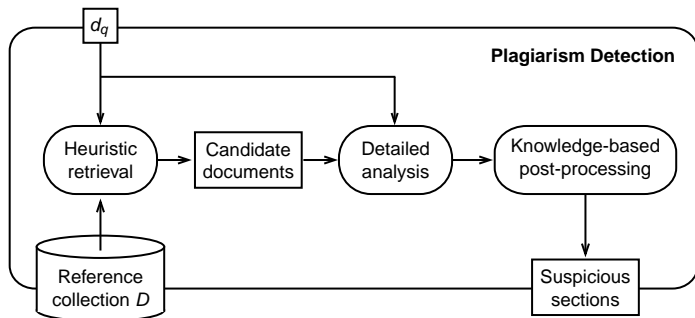
Language and Literature  
Research Center

**NLEL**

Support Language Engineering Life

# PDO: External Plagiarism Detection

## Prototypical Process



# Outline

Introduction

Plagiarism Detection Overview

Preparing a Competition on Plagiarism Detection

First International Competition on Plagiarism Detection

Second International Competition on Plagiarism Detection

And beyond



Language Learning Technology  
**NLEL**  
Support Learning Engineering Life

# PAN-PC-09: Corpus of *Synthetic* Plagiarism

- Plagiarism implies an ethical issue



# PAN-PC-09: Corpus of *Synthetic* Plagiarism

- Plagiarism implies an ethical issue
- Nobody would like to be included in a corpus containing plagiarism!



# PAN-PC-09: Corpus of *Synthetic* Plagiarism

- Plagiarism implies an ethical issue
- Nobody would like to be included in a corpus containing plagiarism!
- Properly anonymising actual cases of plagiarism is a hard task



# PAN-PC-09: Corpus of *Synthetic* Plagiarism

Base texts Texts from Project Gutenberg (<http://www.gutenberg.org>).

Restrictions As the base text is free of copyright, the resulting corpus does **not** have distribution restrictions.

Cases generation All the cases of text reuse are created automatically.

Proper citation No cases of proper citation are included.



Language Learning Technology  
NLEL  
Support Learning. Empowering Life.

## “A newly developed large-scale corpus of *artificial* plagiarism”

- 41 223 documents
- 94 202 artificial plagiarism cases
- It includes cases for intrinsic and external detection methods

<http://www.webis.de/research/corpora>

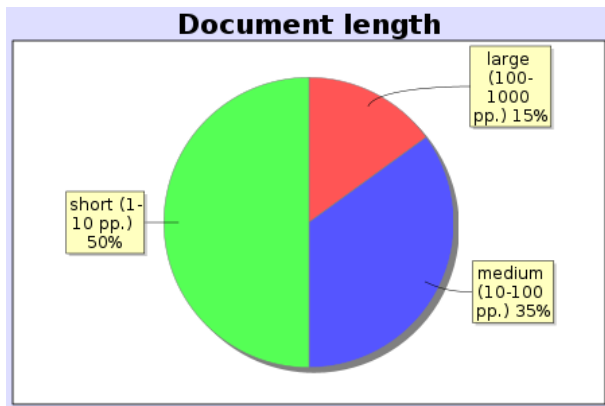


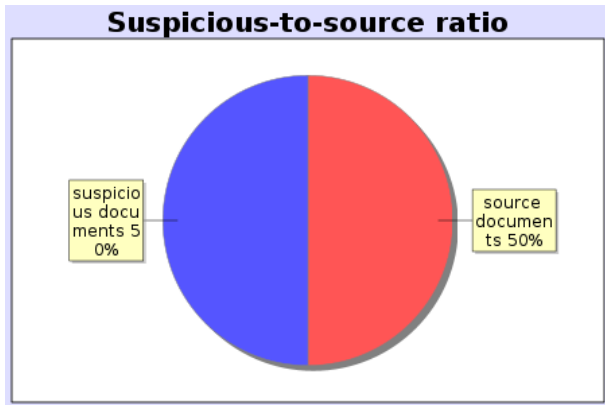
Language Learning Technology  
**NLEL**  
Support Learning. Empowering Life.

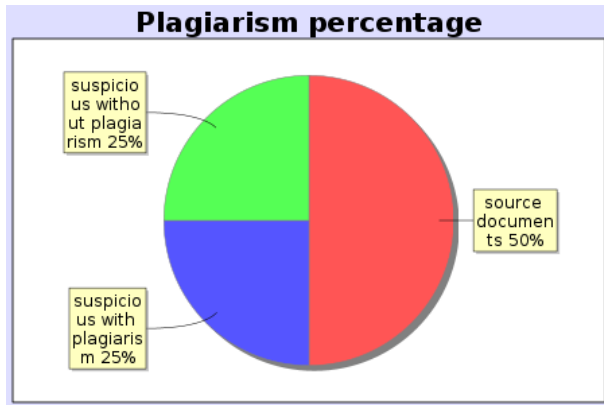
# PAN-PC-09: parameters

- Document length
- Suspicious-to-source ratio
- Plagiarism percentage
- Cases length
- Plagiarism language
- Cases obfuscation

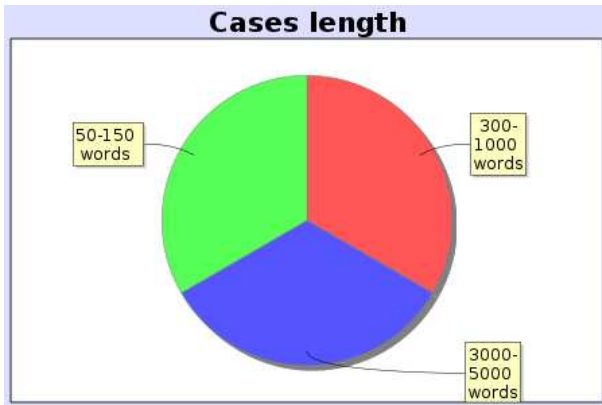


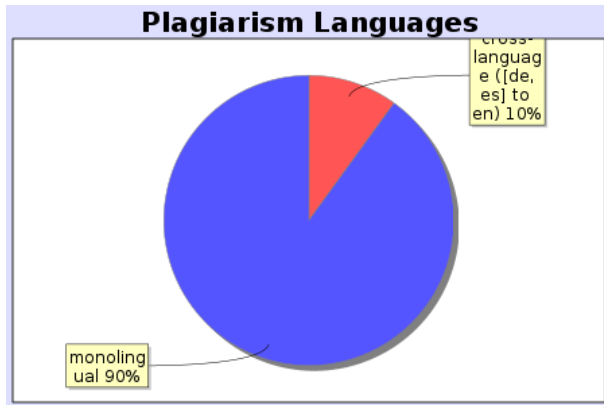






# PAN-PC-09: parameters





## Cases Obfuscation

Paraphrasing, summarization, etc. is simulated by...

- shuffling, removing, inserting short phrases
- replacing semantically related words
- POS preserving shuffling



# Evaluation Measures

We are interested in evaluating the following three main factors:

- 1 plagiarised and —if available— source fragments are retrieved;
- 2 original text fragments are not reported as plagiarised; and
- 3 plagiarised fragments are not detected over and over again.



# Evaluation Measures

- No standard evaluation measures have been previously defined



# Evaluation Measures

- No standard evaluation measures have been previously defined
- Evaluations use to be incomparable and often not even reproducible



# Evaluation Measures

- No standard evaluation measures have been previously defined
- Evaluations use to be incomparable and often not even reproducible
- Properly anonymising actual cases of plagiarism is a hard task



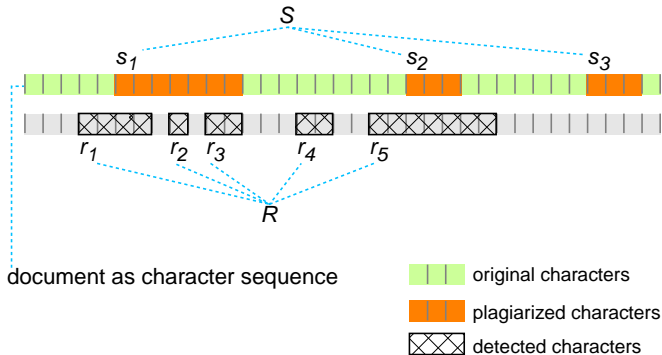
## Precision and Recall

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

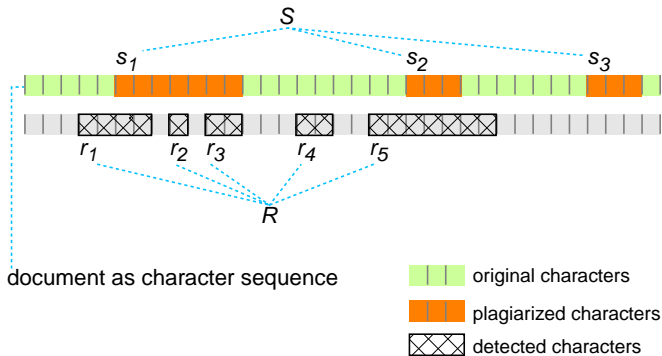
$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$



# Evaluation Measures



# Evaluation Measures

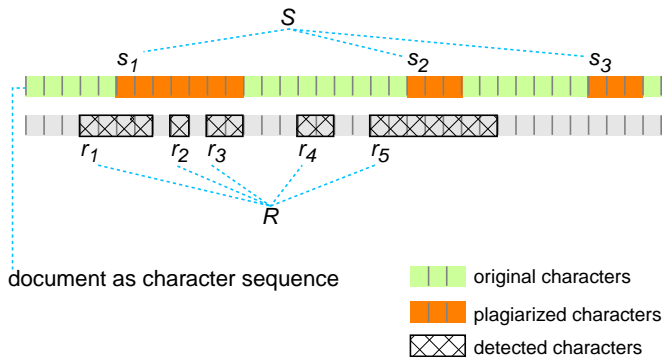


$$rec_{PDA}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|s \cap \bigcup_{r \in R} r|}{|s|}$$

( $\cap$  computes the positionally overlapping characters)



# Evaluation Measures

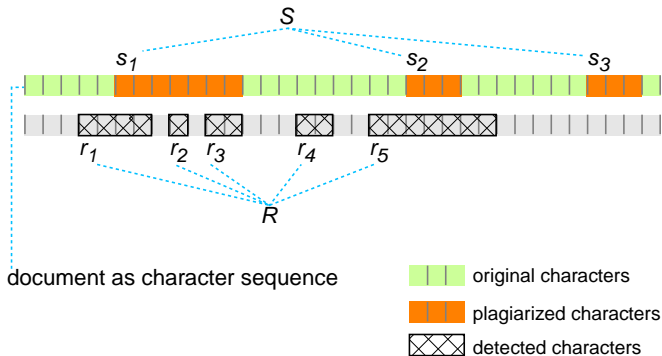


$$prec_{PDA}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|r \cap \bigcup_{s \in S} s|}{|r|}$$

( $\cap$  computes the positionally overlapping characters)



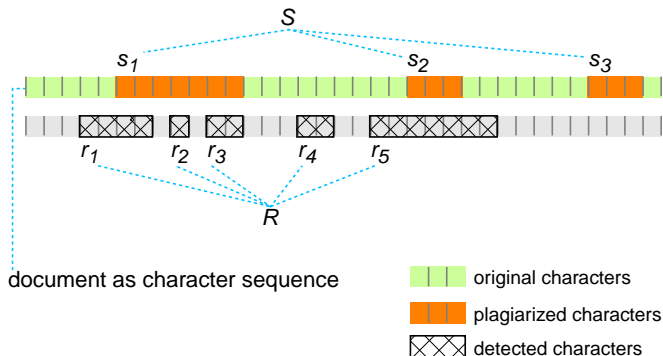
# Evaluation Measures



$$F(S, R) = 2 \cdot \frac{prec(S, R) \cdot rec(S, R)}{prec(S, R) + rec(S, R)}$$



# Evaluation Measures



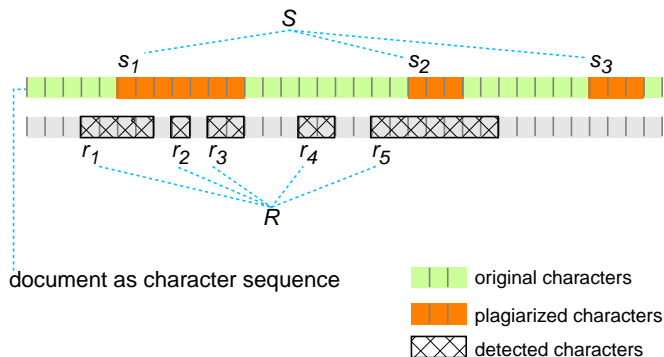
$$gran_{PDA}(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |C_s| \in [1, |R|]$$

$$C_s = \{r \mid r \in R \wedge s \cap r \neq \emptyset\}$$

$$S_R = \{s \mid s \in S \wedge \exists r \in R : s \cap r \neq \emptyset\}$$



# Evaluation Measures



$$overall_{PDA}(S, R) = \frac{F}{\log_2(1 + gran_{PDA})}$$



# Outline

Introduction

Plagiarism Detection Overview

Preparing a Competition on Plagiarism Detection

First International Competition on Plagiarism Detection

Second International Competition on Plagiarism Detection

And beyond



Language Learning Technology  
**NLEL**  
Support Learning Engineering Life

# 1st Competition: Chronology

Mach 2009 Participants were provided with the developing section of the corpus (with annotated cases).



Language Learning Technology  
Research Network  
**NLEL**  
Support Language Engineering Life

# 1st Competition: Chronology

Mach 2009 Participants were provided with the developing section of the corpus (with annotated cases).

May 2009 Test corpus provided (without any annotation).



Language Learning Technology  
Research Centre  
**NLEL**  
Support Language Engineering Life

# 1st Competition: Chronology

- Mach 2009 Participants were provided with the developing section of the corpus (with annotated cases).
- May 2009 Test corpus provided (without any annotation).
- June 2009 Participants submitted their detections to be evaluated.



# 1st Competition: Game rules

**Eligibility** The contest was open to any party planning to attend the PAN competition. No feedback at the time of submission was provided.

**Integrity** The exploitation of potential flaws in the competition corpus to gain advantages was prohibited.

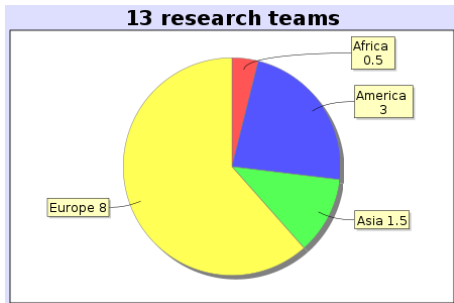
**Text resources** No other text than the one provided in the corpus could be used.

**Winner Selection** One winner of the "External Plagiarism Detection" task, one winner of the "Intrinsic Plagiarism Detection" task, and one overall winner were proclaimed.

**Award** The overall winner was awarded a prize, sponsored by Yahoo! Research.



# 1st Competition: Overview



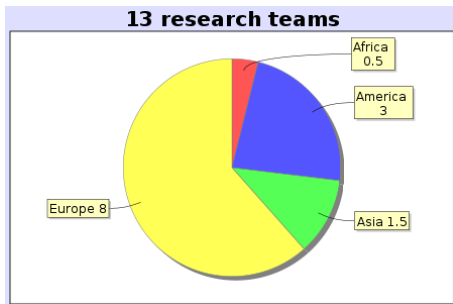
<http://www.webis.de/research/workshopseries/pan-09/competition.html>

<http://ceur-ws.org/Vol-502>



Language Research Group  
**NLEL**  
Support Language Engineering Life

# 1st Competition: Overview



## **Intrinsic Approaches (4 teams)**

Participant	Analyzed features
Stamatatos	character $n$ -grams
Zechner, Muhr, Kern, Granitzer	word freq. class + text frequencies
Seaward, Matwin	Kolmogorov complexity measures

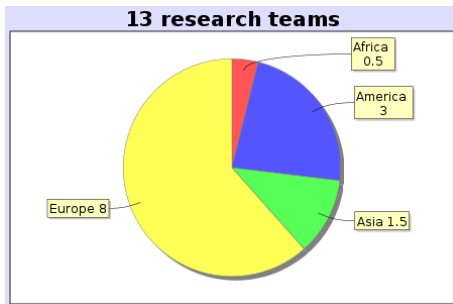
<http://www.webis.de/research/workshopseries/pan-09/competition.html>

<http://ceur-ws.org/Vol-502>



NLEL  
National Laboratory for Intelligent Systems and Applications

# 1st Competition: Overview



## External Approaches (10 teams)

Participant	Comparison units
Grozea, Gehl, Popescu	character $n$ -grams
Kasprzak, Brandejs, Kripac	word $n$ -grams
Basile, Benedetto, Caglioti, Degli Esposti	length $n$ -grams

<http://www.webis.de/research/workshopseries/pan-09/competition.html>



NLEL  
National Laboratory for Language Engineering

# Outline

Introduction

Plagiarism Detection Overview

Preparing a Competition on Plagiarism Detection

First International Competition on Plagiarism Detection

Second International Competition on Plagiarism Detection

And beyond



Language Learning Technology  
**NLEL**  
Support Learning Engineering Life

# 2nd Competition

PAN 2010 LAB

Uncovering Plagiarism, Authorship, and Social Software Misuse

sponsored by



held in conjunction with



- PAN-PC-09 corpus → PAN 2010 training corpus
- PAN 2010 test corpus composed of around 40,000 documents



Language Learning through  
Research and Innovation  
**NLEL**  
Support Learning, Empowering Life



# 2nd Competition

PAN 2010 LAB

Uncovering Plagiarism, Authorship, and Social Software Misuse

sponsored by



held in conjunction with



- PAN-PC-09 corpus → PAN 2010 training corpus
- PAN 2010 test corpus composed of around 40,000 documents
- Cases of “unaware” text reuse are minimised



Language Learning through  
Research and Innovation  
**NLEL**  
Support Learning, Empowering Life



# 2nd Competition

## PAN 2010 LAB

Uncovering Plagiarism, Authorship, and Social Software Misuse

sponsored by



held in conjunction with



- PAN-PC-09 corpus → PAN 2010 training corpus
- PAN 2010 test corpus composed of around 40,000 documents
- Cases of “unaware” text reuse are minimised
- Better obfuscation models



Language Learning through  
Research and Innovation  
**NLEL**  
Support Learning, Empowering Life



# 2nd Competition

PAN 2010 LAB

Uncovering Plagiarism, Authorship, and Social Software Misuse

sponsored by



held in conjunction with



- PAN-PC-09 corpus → PAN 2010 training corpus
- PAN 2010 test corpus composed of around 40,000 documents
- Cases of “unaware” text reuse are minimised
- Better obfuscation models
- Higher relation between a plagiarism and its context



Language and Learning  
Research and Innovation  
NLEL  
Support Learning, Empowering Life



# 2nd Competition

PAN 2010 LAB

Uncovering Plagiarism, Authorship, and Social Software Misuse

sponsored by



held in conjunction with



- PAN-PC-09 corpus → PAN 2010 training corpus
- PAN 2010 test corpus composed of around 40,000 documents
- Cases of “unaware” text reuse are minimised
- Better obfuscation models
- Higher relation between a plagiarism and its context
- No distinction between intrinsic and external cases



Language Learning Technology  
Research Center  
**NLEL**  
Support Learning, Empowering Life

# 2nd Competition

## PAN 2010 LAB

Uncovering Plagiarism, Authorship, and Social Software Misuse

sponsored by



held in conjunction with



- PAN-PC-09 corpus → PAN 2010 training corpus
- PAN 2010 test corpus composed of around 40,000 documents
- Cases of “unaware” text reuse are minimised
- Better obfuscation models
- Higher relation between a plagiarism and its context
- No distinction between intrinsic and external cases
- New humanmade cases using Mechanical Turk



Language Learning through  
Research and Innovation  
Support Learning Engineering Life

NLEL

# 2nd Competition: competitors (~20 teams)

PAN 2010 LAB

Uncovering Plagiarism, Authorship, and Social Software Misuse

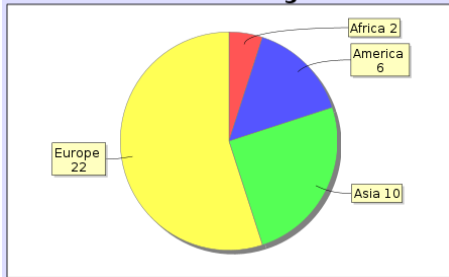
sponsored by

**YAHOO!**  
RESEARCH

held in conjunction with



## 40 researchers registered



<http://pan.webis.de>



Language and Learning Technology  
Research Center  
**NLEL**  
Support Language Engineering Life

# Outline

Introduction

Plagiarism Detection Overview

Preparing a Competition on Plagiarism Detection

First International Competition on Plagiarism Detection

Second International Competition on Plagiarism Detection

And beyond



Language Learning Technology  
Research and Innovation Center  
**NLEL**  
Support Learning Engineering Life

# Cross-Language Plagiarism Detection



Language Engineering  
Research Centre  
**NLEL**  
Support Language Engineering Life

# Cross-Language Plagiarism Detection

*The Party of European Socialists (PES) is a European political party comprising thirty-two socialist, social democratic and labour parties from each European Union member state and Norway.*

*El Partido Socialista Europeo (PSE) es un partido político pan-europeo cuyos miembros son de partidos socialdemócratas, socialistas y laboristas de estados miembros de la Unión Europea, así como de Noruega.*

*Europako Alderdi Sozialista Europar Batasuneko herrialdeetako eta Norvegiako hogeita hamahiru alderdi sozialista, sozialdemokrata eta laborista biltzen dituen alderdia da.*

[Wikipedia, 2010]



Language Learning Technology  
NLEL  
Support Learning. Empowering Life.

# Cross-Language Plagiarism Detection

## WIKIPEDIA

### English

*The Free Encyclopedia*

3 321 000+ articles

### 日本語

フリー百科事典

682 000+ 記事

### Deutsch

*Die freie Enzyklopädie*

1 080 000+ Artikel

### Français

*L'encyclopédie libre*

958 000+ articles

### Italiano

*L'enciclopedia libera*

697 000+ voci

### Português

*A enciclopédia livre*

585 000+ artigos

### Nederlands

*De vrije encyclopedie*

606 000+ artikelen

### Español

*La enciclopedia libre*

608 000+ artículos

### Polski

*Wolna encyklopedia*

706 000+ haseł

### Русский

*Свободная энциклопедия*

547 000+ статей



Language Learning and Education  
Network of Language Learning and Education  
NLEL  
Network of Language Learning and Education

# Cross-Language Plagiarism Detection

- In the 1st competition no team tried to detect the CL plagiarism
- In fact, this is a lack in automatic plagiarism detection nowadays



# Cross-Language Plagiarism Detection

- In the 1st competition no team tried to detect the CL plagiarism
- In fact, this is a lack in automatic plagiarism detection nowadays

[Pouliquen et al., 2003]    Exploit multilingual thesauri in order to detect documents translations

[Potthast et al., 2010]    CL-ESA intends to estimate how semantically similar two texts (in different languages are). It exploits Wikipedia.

[Pinto et al., 2009]    CL-ASA is based on statistical machine translation. It estimates the likelihood of two texts of being valid translations of each other.



Thank you!

<http://pan.webis.de>

[pan@webis.de](mailto:pan@webis.de)

[lbarron@dsic.upv.es](mailto:lbarron@dsic.upv.es)

This research is partially funded by CONACYT-Mexico and the MICINN project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).



Language and Learning Technology  
**NLEL**  
Support Language Engineering Life



# References I



Bernstein, Y. and Zobel, J. (2004).

## A Scalable System for Identifying Co-Derivative Documents.

In *Proceedings of the Symposium on String Processing and Information Retrieval*, pages 55–67. Springer.



Broder, A. (1997).

## On the Resemblance and Containment of Documents.

In *Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29. IEEE Computer Society.



Maurer, H., Kappe, F., and Zaka, B. (2006).

## Plagiarism - A Survey.

*Journal of Universal Computer Science*, 12(8):1050–1084.



Meyer zu Eißlen, S. and Stein, B. (2006).

## Intrinsic plagiarism detection.

*Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research (ECIR 2006)*, LNCS (3936):565–569.



Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., and Rosso, P. (2009).

## A Statistical Approach to Crosslingual Natural Language Tasks.

*Journal of Algorithms*, 64(1):51–60.



Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2010).

## Cross-Language Plagiarism Detection.

*Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*.



Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., and Rosso, P. (2009).

## Overview of the 1st International Competition on Plagiarism Detection.

In [Stein et al., 2009], pages 1–9.



Language Learning and Intelligent Systems  
**NLEL**  
Natural Language Engineering Lab

# References II



Potthast, et al. (2010).

## **An Evaluation Framework for Plagiarism Detection.**

In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.



Pouliquen, B., Steinberger, R., and Ignat, C. (2003).

## **Automatic Identification of Document Translations in Large Multilingual Document Collections.**

In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 401–408, Borovets, Bulgaria.



Schleimer, S., Wilkerson, D., and Aiken, A. (2003).

## **Winnowing: Local Algorithms for Document Fingerprinting.**

In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, New York, NY. ACM.



Stamatatos, E. (2009).

## **Intrinsic Plagiarism Detection Using Character $n$ -gram Profiles.**

In [Stein et al., 2009], pages 38–46.



Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E., editors (2009).

*SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, San Sebastian, Spain. CEUS-WS.org.



Wikipedia (2010).

Party of European Socialists | Partido Socialista Europeo | Europako Alderdi Sozialista .

[Online; accessed 10-February-2010].



Language and Learning Technology  
NLEL  
Support Language Engineering Life