Detection of Cross-Language Text Reuse

Alberto Barrón-Cedeño

Natural Language Engineering Lab, ELiRF
Department of Information Systems and Computation
Universidad Politécnica de Valencia

WIQE 2010
13th September, 2010

Outline

Introduction

- $\mathcal{A}$ Copying words or ideas from someone else without giving credit
- $\mathcal{A}'_1$ Copying the words and ideas from someone else's text without giving credit
- $\mathcal{A}'_2$ Changing words but copying the sentence structure of a source without giving credit
- $\mathcal{A}'_3$ Copiar las palabras o ideas de alguien más sin darle crédito

Introduction

- $\mathcal{A}$ Copying words or ideas from someone else without giving credit.
- $\mathcal{A}'_1$ Copying the words and ideas from someone else's text without giving credit.
- $\mathcal{A}'_2$ Changing words but copying the sentence structure of a source without giving credit.
- $\mathcal{A}'_3$ Copiar las palabras o ideas de alguien más sin darle crédito

$\mathcal{A}'_1$ is plagiarised. $\mathcal{A}'_2$ is not. $\mathcal{A}'_3$ is **cross-language plagiarism**

Introduction

A Beautiful Plagiarism Definition

to take the though or style of another writer whom one has never, never read.

[Bierce, 1911] (Devil's Dictionary)

Introduction

Why CROSS-LANGUAGE plagiarism detection is interesting?

- Plagiarism does not end at language boundaries

Introduction

Why CROSS-LANGUAGE plagiarism detection is interesting?

- Plagiarism does not end at language boundaries
- Authors writing material in different languages are common

Introduction

Why CROSS-LANGUAGE plagiarism detection is interesting?

- Plagiarism does not end at language boundaries
- Authors writing material in different languages are common
- Foreign people (students) feel safe of plagiarising from sources written in their native language

Introduction

Why CROSS-LANGUAGE plagiarism detection is interesting?

- Plagiarism does not end at language boundaries
- Authors writing material in different languages are common
- Foreign people (students) feel safe of plagiarising from sources written in their native language
- **The lack of texts in less resourced languages become cl-plagiarism common**

Why CROSS-LANGUAGE plagiarism detection is interesting?

(www.wikipedia.org; consulted: 13th July, 2010)

Outline

SIGIR Doctoral Consortium

- plagiarism detection
- the cross-language case and how we are trying to approach it
- the PAN competition (corpus + evaluation)

Alberto Barrón-Cedeño. On the Mono- and Cross-Language Detection of Text Reuse and Plagiarism. SIGIR 2010 (Doctoral Consortium)

SIGIR Doctoral Consortium

- plagiarism detection
- the cross-language case and how we are trying to approach it
- the PAN competition (corpus + evaluation)

Questions:

Collections generation

- How to get a freely available corpus of actual plagiarism?
- How valid is simulated plagiarism generated by crowdsourcing?

Alberto Barrón-Cedeño. On the Mono- and Cross-Language Detection of Text Reuse and Plagiarism. SIGIR 2010 (Doctoral Consortium)

SIGIR Doctoral Consortium

- plagiarism detection
- the cross-language case and how we are trying to approach it
- the PAN competition (corpus + evaluation)

Questions:

Collections generation

- How to get a freely available corpus of actual plagiarism?
- How valid is simulated plagiarism generated by crowdsourcing?

CL detection

- Is there a general bilingual statistical dictionary available? Otherwise, how to build it?
- CL text reuse in Wikipedia and newspapers is an option?

Alberto Barrón-Cedeño. On the Mono- and Cross-Language Detection of Text Reuse and Plagiarism. SIGIR 2010 (Doctoral Consortium)

SIGIR Doctoral Consortium

Doug Oard
Alistair Moffat
Jamie Callan

Paul Thomas
Maarten de Rijke
Mounia Lalmas

-
-
-
-

SIGIR Doctoral Consortium

Doug Oard
Alistair Moffat
Jamie Callan
Paul Thomas
Marteen de Rijke
Mounia Lalmas

- Stop naming it plagiarism → **text reuse**
-
-
-

SIGIR Doctoral Consortium

Doug Oard
Alistair Moffat
Jamie Callan
Paul Thomas
Marteen de Rijke
Mounia Lalmas

- Stop naming it plagiarism → **text reuse**
- CL is too complicated. Instead, think about a comparison of models and speed them up
-
-

SIGIR Doctoral Consortium

Doug Oard
Alistair Moffat
Jamie Callan
Paul Thomas
Marteen de Rijke
Mounia Lalmas

- Stop naming it plagiarism → **text reuse**
- CL is too complicated. Instead, think about a comparison of models and speed them up
- Use the corpus XX and evaluate Precision
-

SIGIR Doctoral Consortium

Doug Oard
Alistair Moffat
Jamie Callan
Paul Thomas
Marteen de Rijke
Mounia Lalmas

- Stop naming it plagiarism → **text reuse**
- CL is too complicated. Instead, think about a comparison of models and speed them up
- Use the corpus XX and evaluate Precision
- Try to contact Paul Clough at U. of Sheffield

SIGIR Doctoral Consortium

Doug Oard
Alistair Moffat
Jamie Callan
Paul Thomas
Marteen de Rijke
Mounia Lalmas

- Stop naming it plagiarism → **text reuse**
- CL is too complicated. Instead, think about a comparison of models and speed them up
- Use the corpus XX and evaluate Precision
- Try to contact Paul Clough at U. of Sheffield

Proposal (*i*) Download a collection of papers (ArXiv, ACL);
(*ii*) search for **reused text**

Detection of Simple Plagiarism in Computer Science Papers

- stressed that *self* -plagiarism is a very important problem nowadays
- he even offered a case study, just to do friends (long story)

[HaCohen-Kerner et al., 2010]

Detection of Simple Plagiarism in Computer Science Papers

- stressed that *self* -plagiarism is a very important problem nowadays
- he even offered a case study, just to do friends (long story)

- $10,000$ documents from ACL
- Comparison of (38) models and options of similarity estimation
- Knowledge based division of documents: abstract, references, first third, etc
- He noted that analysisng abstracts is one of the best options
- Evaluation: runtime + expert analysis

[HaCohen-Kerner et al., 2010]

Outline

CL: Thesaurus based

EUROVOC Thesaurus-based

- Thesaurus catalogued **manually**
- Available in the **18** EU languages

[Pouliquen et al., 2003]

CL: Thesaurus based

EUROVOC Thesaurus-based

- Thesaurus catalogued **manually**
- Available in the **18** EU languages

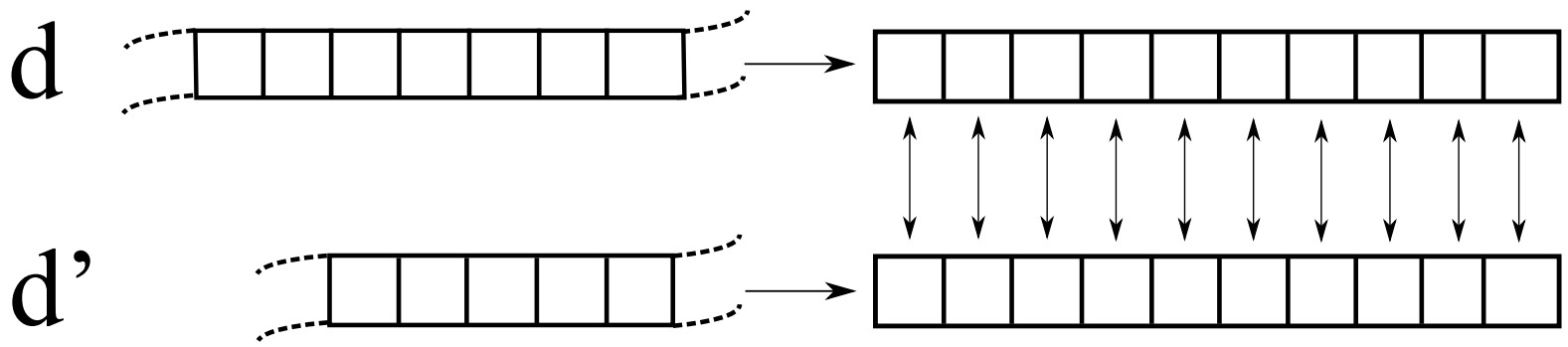
Example *"transport of dangerous goods"* lemmas

Lemma	Weight	Lemma	Weight
dangerous goods	33	radioactive material	19
by road	19	carriage	19
dangerous	18	plutonium	17
radioactive waste	15	nuclear fuel	15
shipment	15	adr	14
bind for	13	tank	13
receptacle	13	transport	13
pollute	12	nuclear waste	12

[Pouliquen et al., 2003]

CL: Thesaurus based

- $d \in L$ and $d' \in L'$ are mapped into a vector of thesaurus descriptor terms

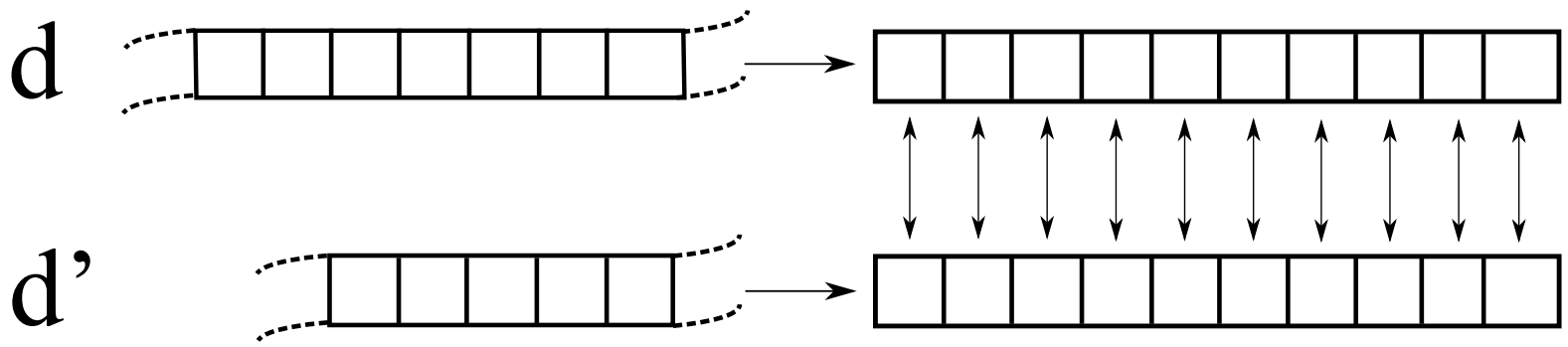


$$sim(d, d') = cos(\theta_{\mathbf{d},\mathbf{d'}})$$

[Pouliquen et al., 2003]

CL: Thesaurus based

- $d \in L$ and $d' \in L'$ are mapped into a vector of thesaurus descriptor terms



$$sim(d, d') = cos(\theta_{\mathbf{d},\mathbf{d'}})$$

- We just obtained it from the EU commission last week!

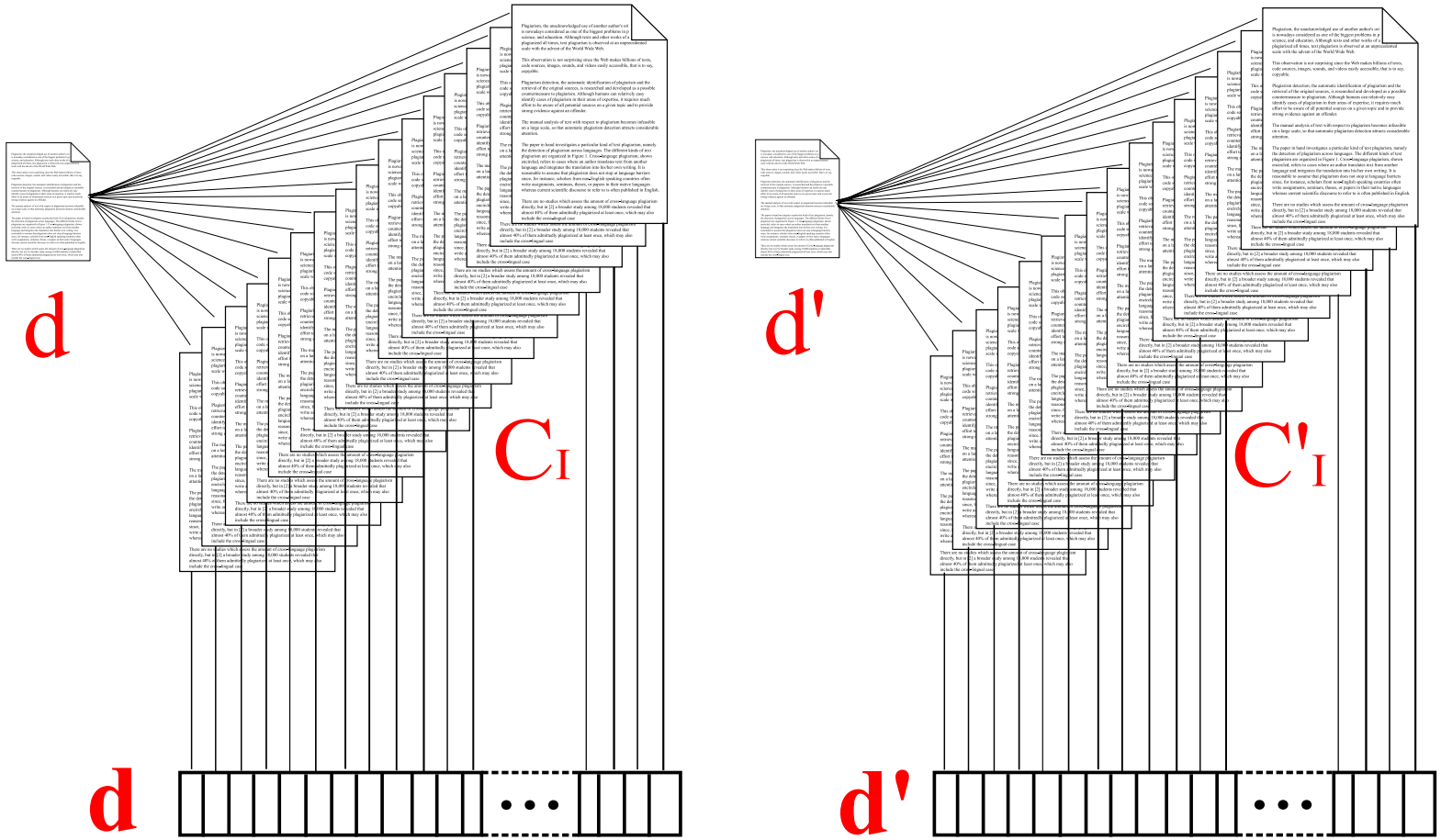
[Pouliquen et al., 2003]

CL: Explicit Semantic Analysis

- A significant comparable corpus C is required
- $d \in L$ ($d' \in L'$) is represented as a vector of relations to the index collection C_I (C'_I)
- The similarities are computed using a monolingual retrieval model such as the VSM
- Wikipedia is one of the biggest comparable corpora nowadays

[Potthast et al., 2008]

CL: Explicit Semantic Analysis



[Potthast et al., 2008]

CL: Alignment-based Similarity Analysis

- How likely is that d is a valid translation of d'?
- A two-step probabilistic translation and similarity analysis
- An adaptation of basic principles statistical MT

[Pinto et al., 2009]

CL: Alignment-based Similarity Analysis

Baye's rule for statistical Machine Translation:

$$p(d' \mid d_q) = \frac{p(d')\ p(d_q \mid d')}{p(d_q)}$$

- $p(d_q)$ does not depend on d' and is therefore neglected
- $p(d_q \mid d')$ is a *translation model probability* (statistical bilingual dictionary)
- . $p(d')$ is the *language model probability*

[Brown et al., 1993]

CL: Alignment-based Similarity Analysis

$$p(d' \mid d_q) = p(d')\ p(d_q \mid d')$$

We propose two adaptations:

- The adapted translation model is a non-probabilistic measure $w(d_q \mid d')$
- The language model is replaced by a *length model* $\varrho(d')$ that depends on document length

$$\varphi(d_q, d') = s(d' \mid d_q) = \varrho(d')\ w(d_q \mid d').$$

[Barrón-Cedeño et al., 2008, Pinto et al., 2009, Potthast et al., 2010]

CL: Alignment-based Similarity Analysis

The translation model depends on a bilingual dictionary (estimated by the IBM M1)

es	**en**	$p(es, en)$
certifica	certifies	0.420329
certifica	certify	0.164481
certifica	certified	0.109649
certifica	certifying	0.091375
certifica	hereby	0.054824
certifica	that	0.050577
certifica	has	0.035947
certifica	declare	0.018275
certifica	licence	0.018271

Translation model

$$p(d \mid d') = \prod_{x \in d} \sum_{y \in d'} p(x, y)$$

Adapted translation model (document level)

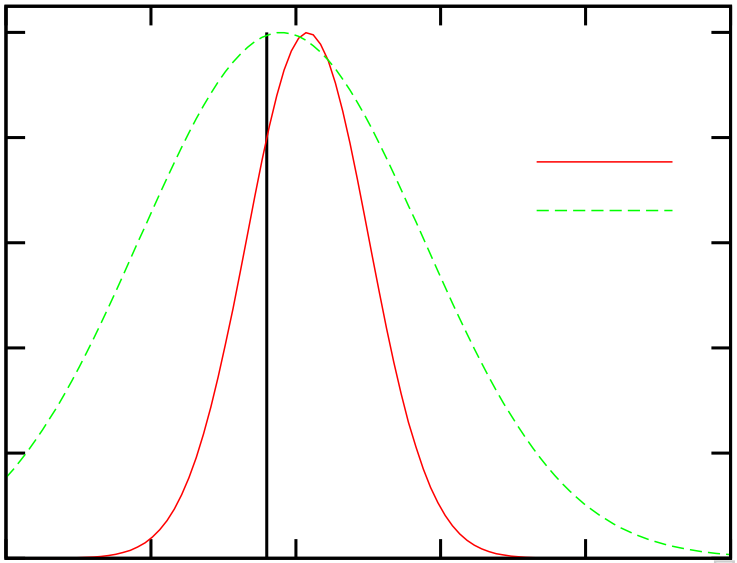
$$w(d \mid d') = \sum_{x \in d} \sum_{y \in d'} p(x, y)$$

- $w(d \mid d')$ increases if valid translations (x, y) appear in the implied vocabularies.
- For a word x, with $p(x, y) = 0$ for all $y \in d'$, $w(d \mid d')$ is decreased by ϵ, in our case $\epsilon = 0.1$.

Length Model

- It is expected that the length of the translation documents d and d' is closely related [Pouliquen et al., 2003]

$$\varrho(d') = e^{-0.5\left(\frac{\frac{|d'|}{|d|}-\mu}{\sigma}\right)^2}$$



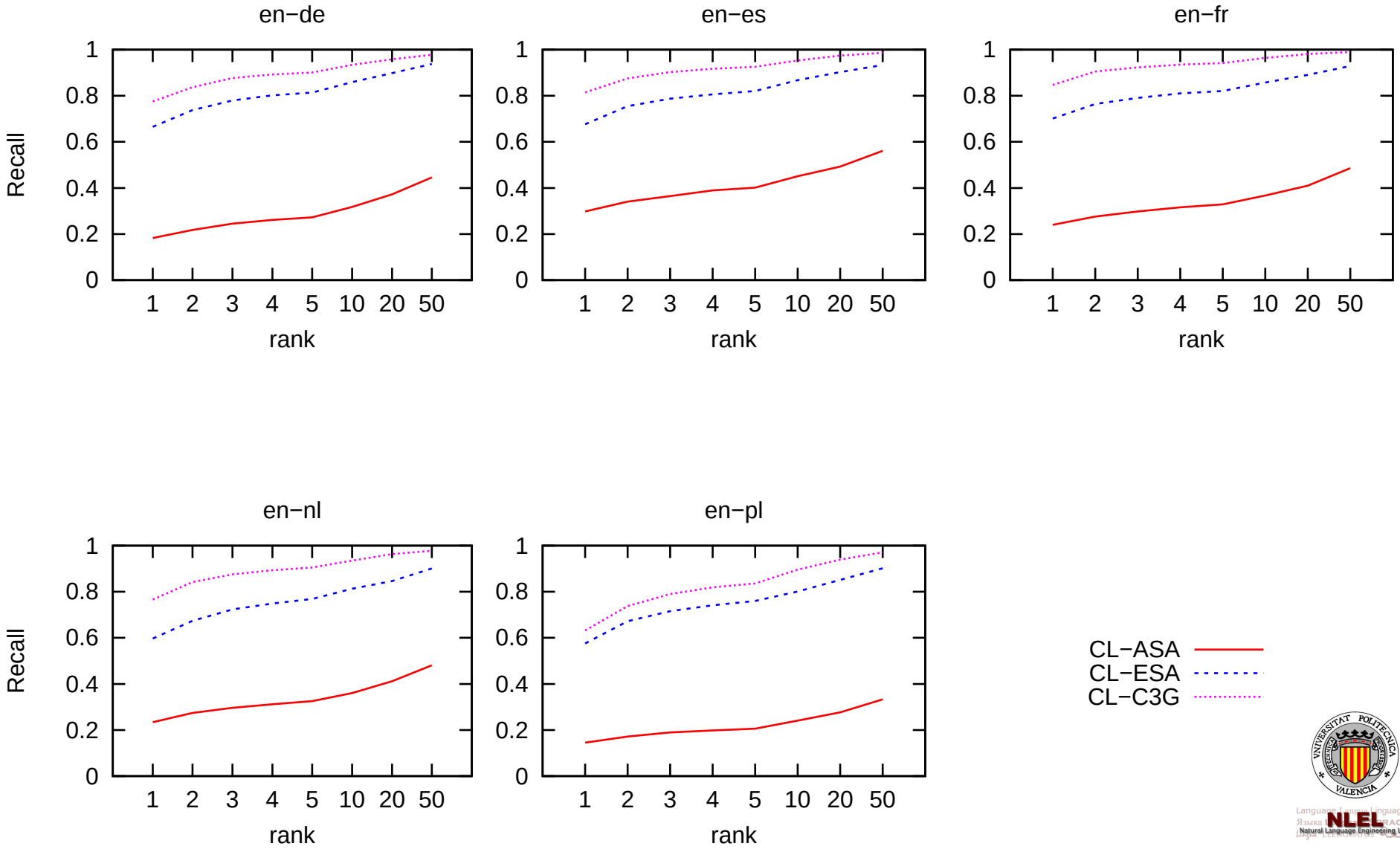
CL: Character n-grams

Character n-grams use to be common languages with syntactical similarities.

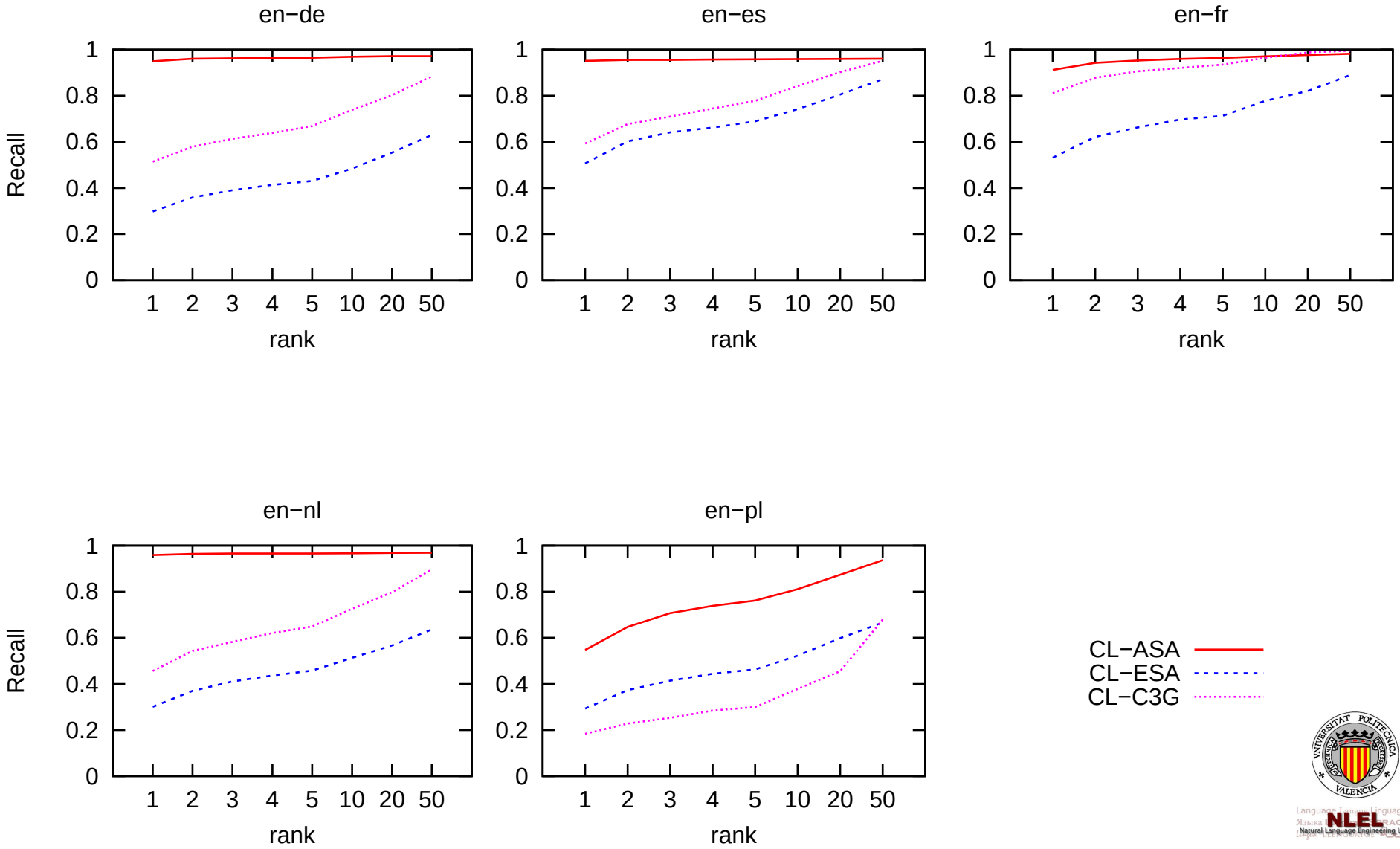
- $\Sigma = \{a, \ldots, z, 0, \ldots, 9\}$,
- $n = 3$
- $tfidf$-weighting
- Cosine similarity

[Mcnamee and Mayfield, 2004]

CL: Cross-language ranking (Wikipedia)



CL: Cross-language ranking (JRC-Acquis)



CL: And for less related languages?

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]

CL: And for less related languages?

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.

The corresponding articles contain around $2,000$, $1,300$, and only 100 words!

[Wikipedia, 2010]

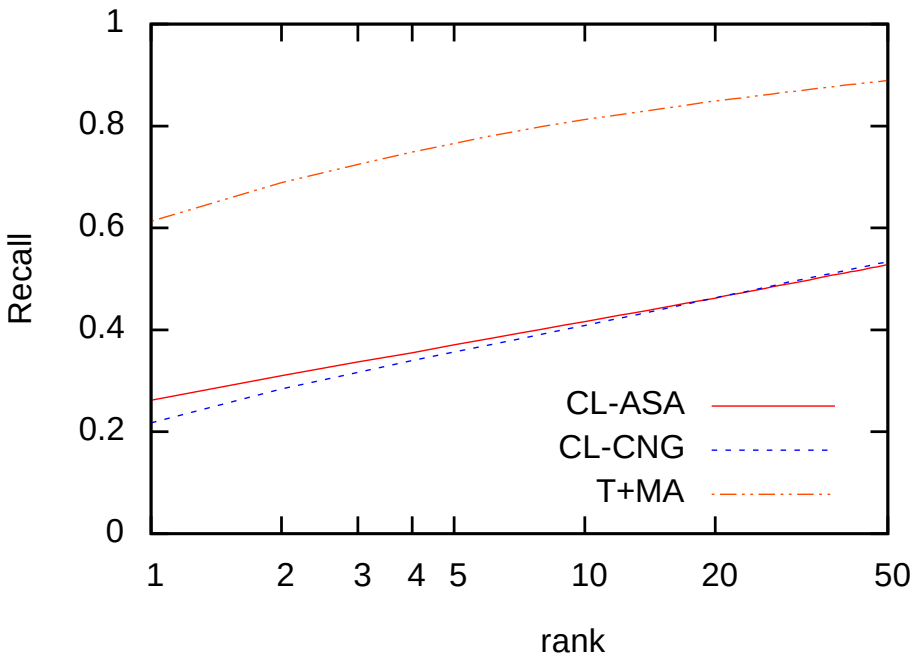
CL: Less Resourced Languages

Framework

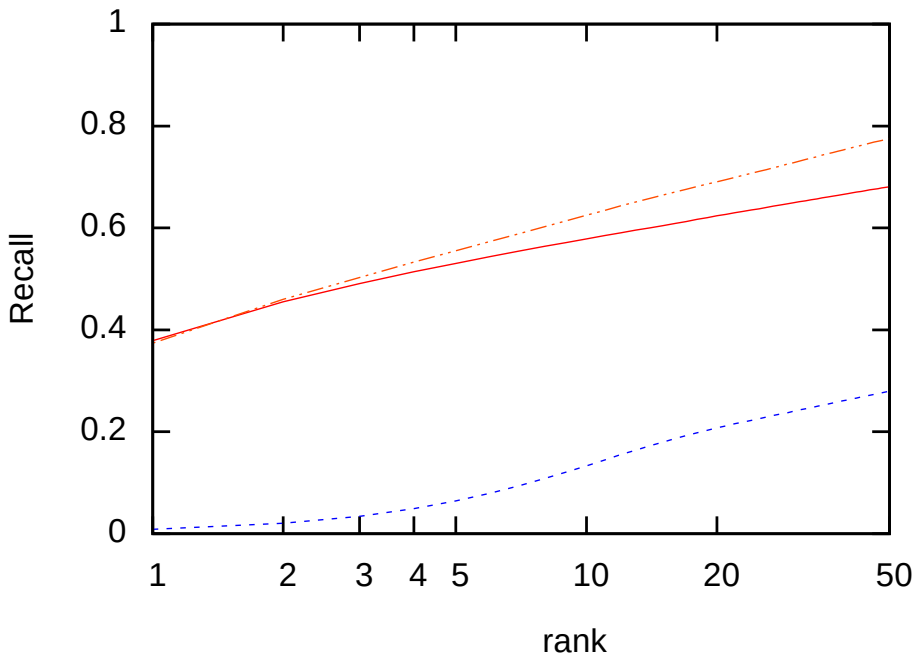
- Two parallel corpora:
  - software: a translation memory (en-eu)
  - consumer: extracts from a multilingual magazine (es-eu)
- The entire corpus is a “big” document
- We perform sentence level similarity estimation

(corpora provided by Elhuyar Fundazioa and Consumer)

CL: Less Resourced Languages



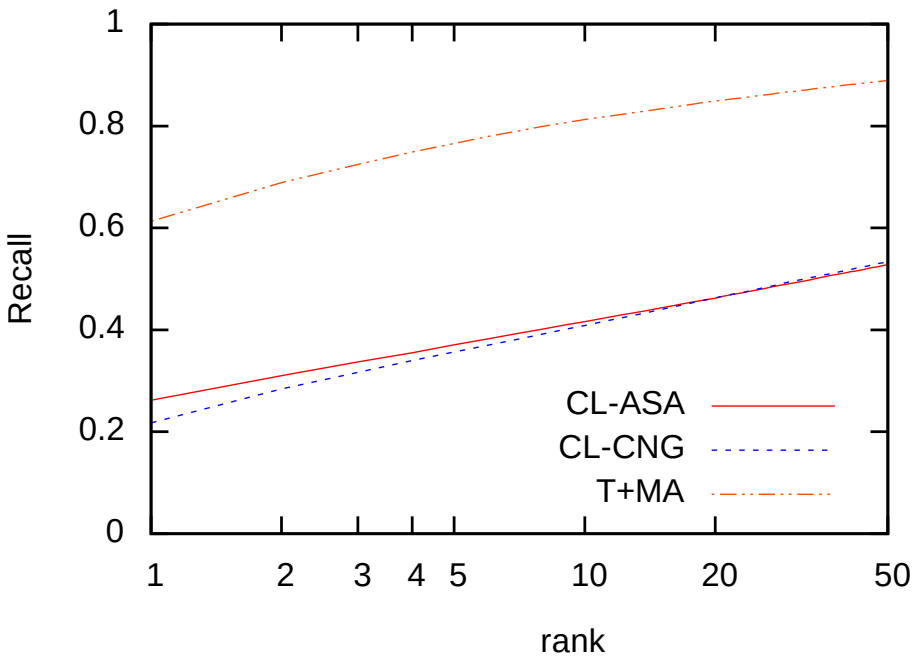
(a) es-eu



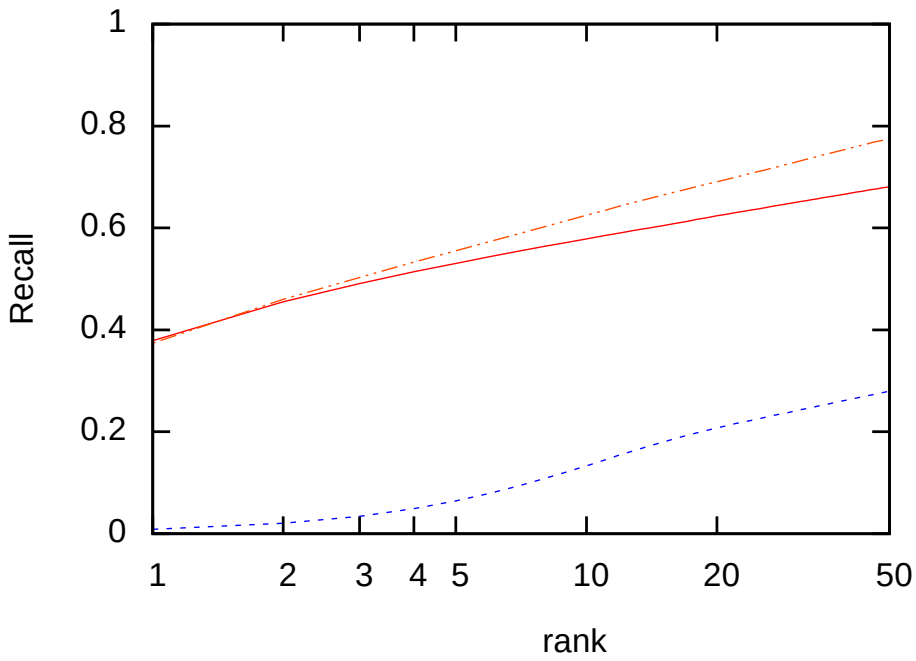
(b) en-eu

[Barrón-Cedeño et al., 2010]

CL: Less Resourced Languages



(c) es-eu

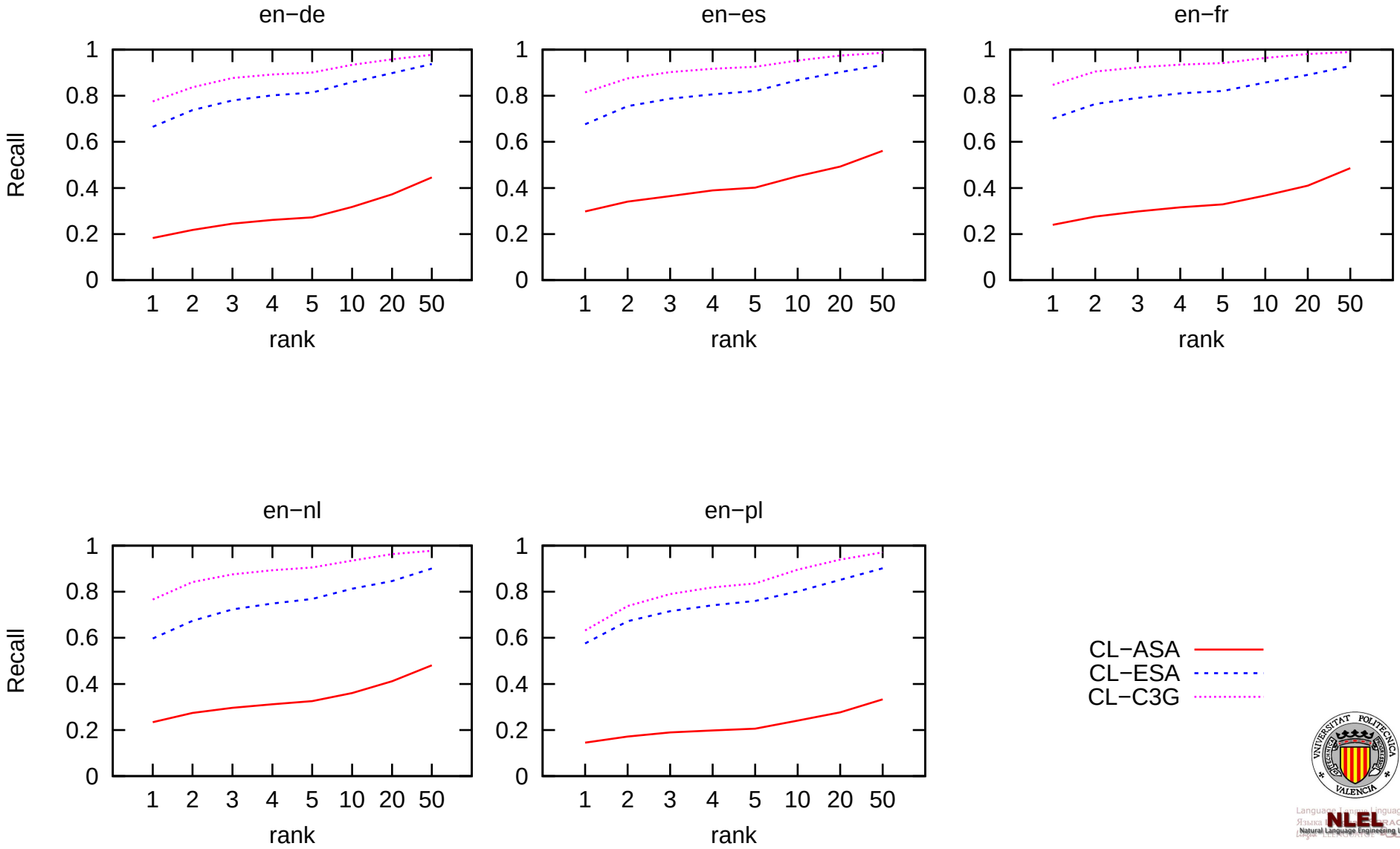


(d) en-eu

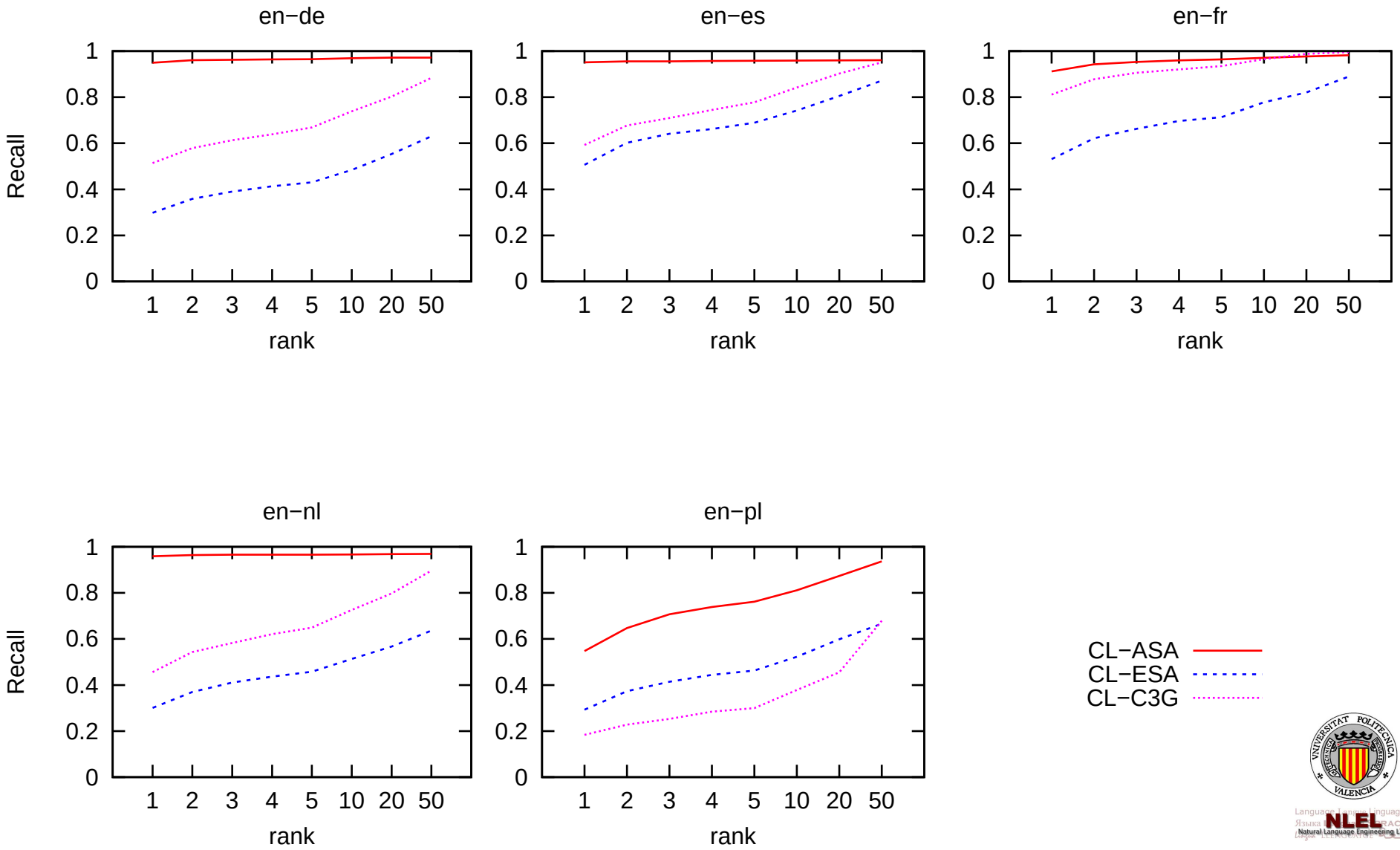
And we are not working with Greek, Arabic, Chinese…!

[Barrón-Cedeño et al., 2010]

CL: Cross-language ranking (Wikipedia)



CL: Cross-language ranking (JRC-Acquis)



Outline

How “comparable” Wikipedia is?

Some authors refer to Wikipedia as one of the biggest comparable corpora at hand.

[Mohammadi and GhasemAghaee, 2010, Potthast et al., 2008].

How “comparable” Wikipedia is?

Some authors refer to Wikipedia as one of the biggest comparable corpora at hand.

[Mohammadi and GhasemAghaee, 2010, Potthast et al., 2008].

A corpus can be considered comparable if:

- it contains the same proportions of texts of the same genres;
- it contains the same proportions of texts in the same domains in a range of different languages; and
- such texts are sampled on the same period.

[McEnery and Xiao, 2007]

How “comparable” Wikipedia is?

Mexico City

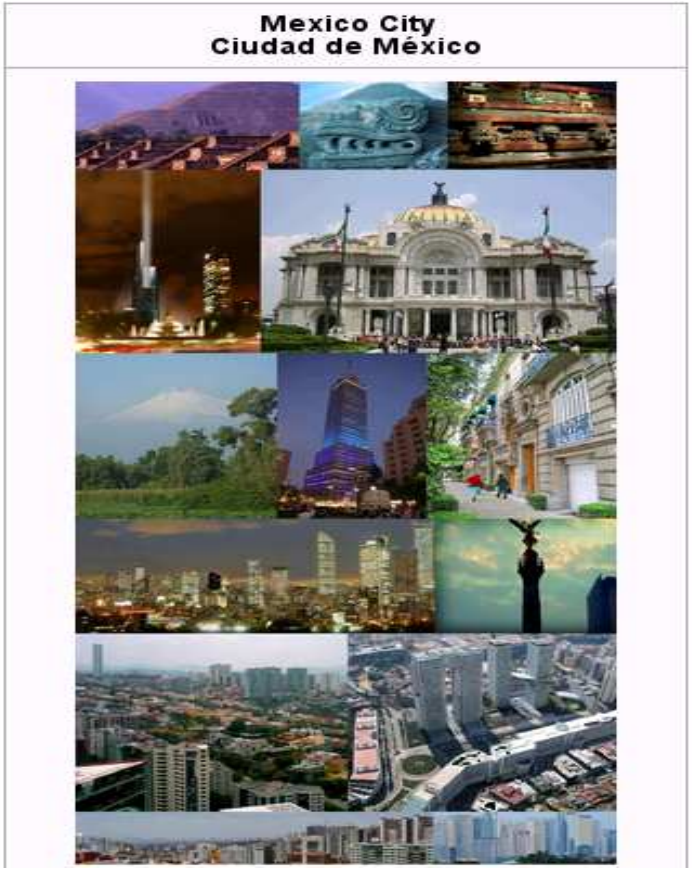
Mexico City (Spanish: *Ciudad de México*) is the capital and largest city in Mexico as well as the largest city in North America and second largest city in the world. Mexico City is also the **Federal District** (Distrito Federal), seat of the federal government [3], a federal entity within Mexico which is not part of any one of the 31 Mexican states but belongs to the federation as a whole. Mexico City is the most important political, cultural, and financial center in the country.

As an "alpha" global city[4] Mexico City is one of the most important financial centers in America[5] which is located in the Valley of Mexico, a large valley in the high plateaus at the center of Mexico, at an altitude of 2,240 metres (7,350 ft). The city consists of sixteen boroughs.

The 2009 estimated population for the city proper exceeds 8.84 million people,[2] and with a land area of 1,485 square kilometres (573 sq mi)[6] According to the most recent definition agreed upon by the federal and state governments, the Mexico City metropolitan area population is 21.2 million people,[2] making it the largest metropolitan area in the Americas and the third largest agglomeration in the world.[7]

Mexico City has a Gross Domestic Product (GDP) of $390 billion USD in 2008, making Mexico City the eighth richest city in the world.[8] The city was responsible for generating 21% of Mexico's Gross Domestic Product and the metropolitan area accounted for 34% of total national GDP.[9] As of 2008, the city proper, as opposed to the metropolitan area, had a nominal income per capita of $25,258 USD, on par with the GDP per capita of Portugal, and significantly above nations such as South Korea, and Czech Republic.[10]

The city was originally built on an island of Lake Texcoco by the Aztecs in 1325 as Tenochtitlan, which was almost completely destroyed in the siege of 1521, and subsequently redesigned and rebuilt in accordance with the Spanish urban standards. In 1524, the municipality of Mexico City was established, known as *México Tenustitlán*.[11] and as of 1585 it was officially known as *La Ciudad de*



How “comparable” Wikipedia is?

Ciudad de México

La **ciudad de México** es el Distrito Federal (abreviado "D. F."), capital de los Estados Unidos Mexicanos y sede de los poderes federales de la Unión[3] y constituye una de las 32 entidades federativas. Comúnmente, en el resto del país, a ésta se le llama de manera abreviada "México" o también "Distrito Federal", mientras que en el extranjero suele denominarse simplemente "ciudad de México".

La ciudad de México es el centro político y económico del país. Su área metropolitana es la novena más poblada del mundo,[4] y la segunda más poblada de Norteamérica.[5] La ciudad de México ocupa el octavo sitio entre las ciudades más ricas del mundo, al tener un PIB de 315.000 millones de dólares que, según se estima, se duplicará para el 2020.

Ocupa una décima parte de la Cuenca de México en el centro-sur del país, que entre otros comprende al Valle de México donde se encuentra el núcleo original y mayor de la ciudad, en un territorio que formó parte de la cuenca lacustre del lago de Texcoco. siendo la ciudad más rica y poblada del país, la zona comprendida dentro del Distrito Federal comprende poco más de ocho millones de habitantes en el 2005,[6] y ocupa el segundo lugar como entidad federativa, solamente detrás del estado de México, que comparte buena parte del área metropolitana de la ciudad. En su crecimiento demográfico, la ciudad de México fue incorporando a numerosos poblados que se encontraban en las cercanías. A mediados del siglo XX, su área metropolitana desbordaba los límites territoriales del Distrito Federal, y se extendía sobre 40 municipios del estado de México y un municipio del estado de Hidalgo, según la definición oficial de la Zona Metropolitana de la Ciudad de México (ZMCM), elaborada en el 2003 por los gobiernos locales, estatales y federal.[7]

La ZMCM estaba habitada en 2005 por 19.331.365 personas, casi el 20 por ciento de la población total del país. De acuerdo con las proyecciones del Consejo Nacional de Población (Conapo); para el 1 de julio de 2007 se estimaba una población de 8.193.899 habitantes para la ciudad, y de 19.704.125 habitantes para toda la Zona Metropolitana. El ingreso per cápita del Distrito Federal ascendía en 2008 a 281.110 pesos mexicanos, lo cual equivalía en dólares nominales de septiembre de 2008 a 25.258 dólares[8] -cifra similar a la de países como la República Checa o Corea del Sur.

Distrito Federal
Ciudad de México

Capital de México

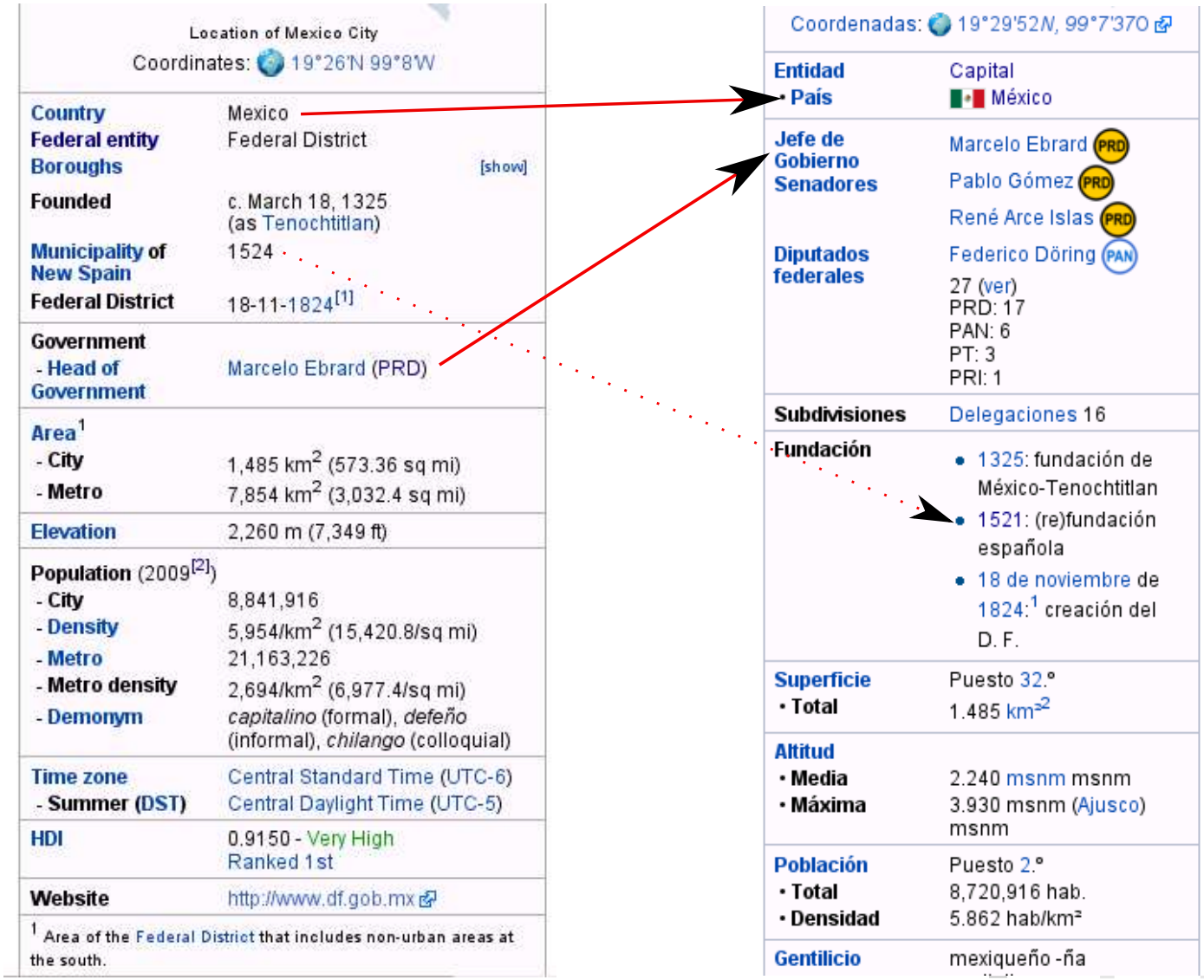
Escudo

Coordenadas: 19°29'52N, 99°7'37O

Entidad • **País**	Capital México
Jefe de Gobierno	Marcelo Ebrard (PRD)
Senadores	Pablo Gómez (PRD)
	René Arce Islas (PRD)
Diputados federales	Federico Döring (PAN)
	27 (ver) PRD: 17

How “comparable” Wikipedia is?

Mexico City / Ciudad de México (infobox)



Location of Mexico City
Coordinates: 19°26′N 99°8′W

Country	Mexico
Federal entity	Federal District
Boroughs	[show]
Founded	c. March 18, 1325 (as Tenochtitlan)
Municipality of New Spain	1524
Federal District	18-11-1824[1]
Government - Head of Government	Marcelo Ebrard (PRD)
Area[1]	
- City	1,485 km² (573.36 sq mi)
- Metro	7,854 km² (3,032.4 sq mi)
Elevation	2,260 m (7,349 ft)
Population (2009[2])	
- City	8,841,916
- Density	5,954/km² (15,420.8/sq mi)
- Metro	21,163,226
- Metro density	2,694/km² (6,977.4/sq mi)
- Demonym	*capitalino* (formal), *defeño* (informal), *chilango* (colloquial)
Time zone	Central Standard Time (UTC-6)
- Summer (DST)	Central Daylight Time (UTC-5)
HDI	0.9150 - Very High Ranked 1st
Website	http://www.df.gob.mx

[1] Area of the Federal District that includes non-urban areas at the south.

Coordenadas: 19°29′52N, 99°7′37O

Entidad • País	Capital México
Jefe de Gobierno Senadores	Marcelo Ebrard (PRD) Pablo Gómez (PRD) René Arce Islas (PRD)
Diputados federales	Federico Döring (PAN) 27 (ver) PRD: 17 PAN: 6 PT: 3 PRI: 1
Subdivisiones	Delegaciones 16
Fundación	• 1325: fundación de México-Tenochtitlan • 1521: (re)fundación española • 18 de noviembre de 1824:[1] creación del D. F.
Superficie • Total	Puesto 32.º 1.485 km²
Altitud • Media • Máxima	 2.240 msnm msnm 3.930 msnm (Ajusco) msnm
Población • Total • Densidad	Puesto 2.º 8,720,916 hab. 5.862 hab/km²
Gentilicio	mexiqueño -ña

How “comparable” Wikipedia is?

Some facts

- Articles on the same topic exist in plenty of languages
- The extent, time, and quality are hardly comparable in the most of the cases.

Reuse of Wikipedia articles

- Reuse of text across related articles
- Reuse of text outside of Wikipedia
- Cross-language text reuse

Cross-Language Reuse in Wikipedia

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]

Cross-Language Reuse in Wikipedia

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.

The corresponding articles contain around $2,000$, $1,300$, and only 100 words!

[Wikipedia, 2010]

Who Reused What?

1. Is this a case of cross-language text reuse?

Who Reused What?

1. Is this a case of cross-language text reuse?
2. What fragment is the source?

- the English one!

Who Reused What?

1. Is this a case of cross-language text reuse?
2. What fragment is the source?

- the English one!
- the longest one

- the first created
- the language “linked” to the topic

Who Reused What?

1. Is this a case of cross-language text reuse?
2. What fragment is the source?

- the English one!
- the longest one

- the first created
- the language “linked” to the topic

11 Nov 07 The English fragment is generated

17 May 09 The English fragment is (slightly) modified

Who Reused What?

1. Is this a case of cross-language text reuse?
2. What fragment is the source?

- the English one!
- the longest one

- the first created
- the language “linked” to the topic

11 Nov 07 The English fragment is generated

17 May 09 The English fragment is (slightly) modified

13 Jun 04 The Spanish fragment is created

19 Jul 07 The Spanish fragment is (slightly) modified

Cross-Language Reuse in Wikipedia

Who Reused What?

1. Is this a case of cross-language text reuse?
2. What fragment is the source?

- the English one!
- the longest one
- the first created
- the language “linked” to the topic

11 Nov 07 The English fragment is generated

17 May 09 The English fragment is (slightly) modified

13 Jun 04 The Spanish fragment is created

19 Jul 07 The Spanish fragment is (slightly) modified

7 Aug 08 The Basque fragment is created (at article creation time)

Research Questions

- are d_L and $d_{L'}$ related?
- are the text fragments $t \in d_L$ and $t' \in d_{L'}$ a translation (reuse) of each other?
- is $t \in d_L$ the source of $t' \in d_{L'}$?

Research Questions

- are d_L and $d_{L'}$ related?
- are the text fragments $t \in d_L$ and $t' \in d_{L'}$ a translation (reuse) of each other?
- is $t \in d_L$ the source of $t' \in d_{L'}$?

Wikimedia Research Questions

- What Wikipedia language editions exist, and why?
- What factors make a language edition grow?
- How are lacks of language planning (e.g. a lack of standardization) dealt with?
- **What kind of inter-Wikipedia collaboration exists?**
- **How to compare language editions to each other?**

http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wikidemia

Articles Alignment

- $d \in L$ and $d' \in L'$ are articles on the same concept c
- alignment of t, t' such that $t \in d$ and $t' \in d'$

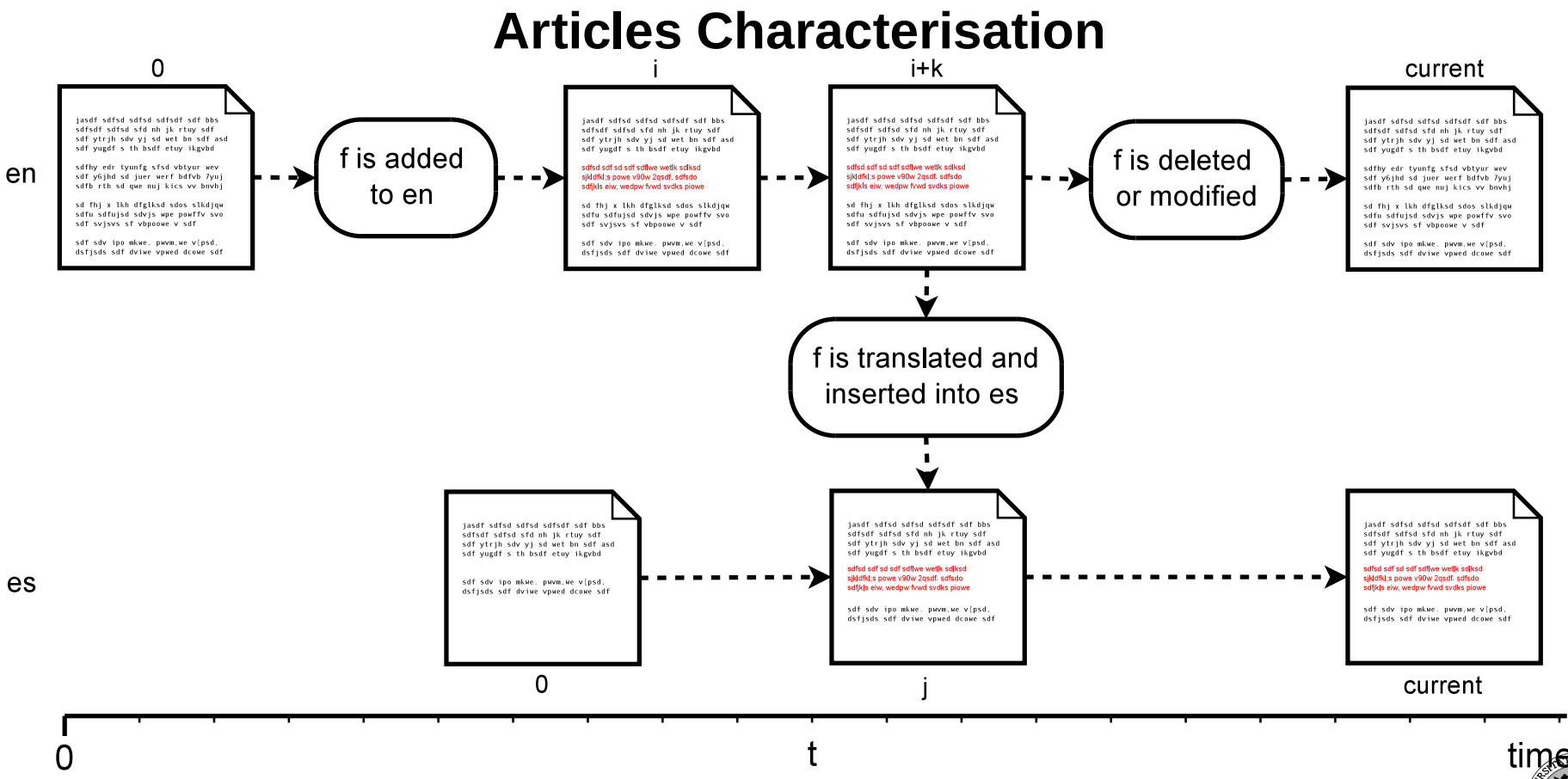
Articles Alignment

- $d \in L$ and $d' \in L'$ are articles on the same concept c
- alignment of t, t' such that $t \in d$ and $t' \in d'$

Text features

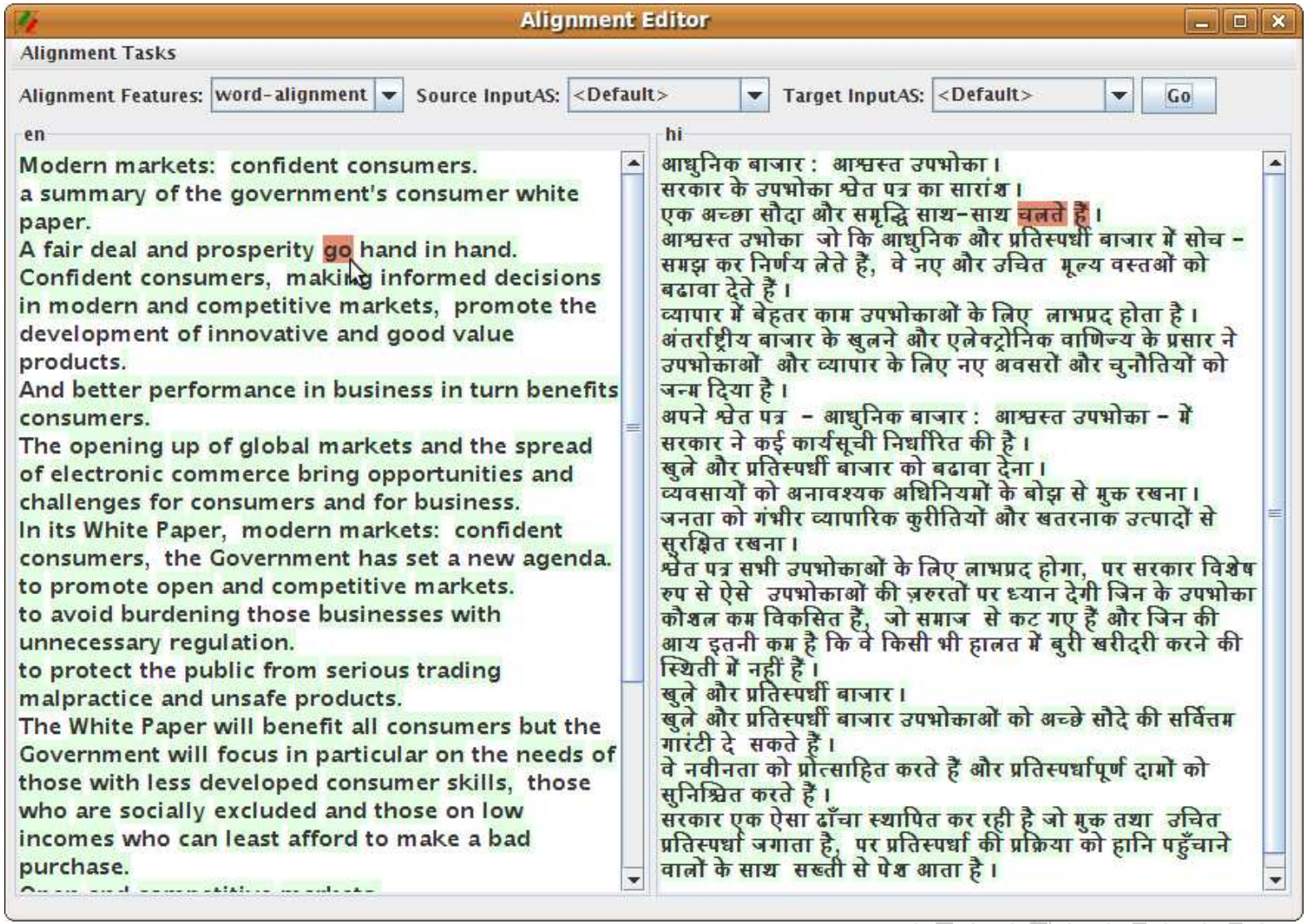
- Named Entities [Friburger and Maurel, 2002, Toral and Muñoz, 2006]
- CL-CNG [Mcnamee and Mayfield, 2004]
- CL-ESA [Potthast et al., 2008]
- CL-ASA [Barrón-Cedeño et al., 2008, Pinto et al., 2009]
- Thesauri-based [Steinberger et al., 2002, Ceska et al., 2008]
- Translation + Monolingual similarity estimation
- Cognates [Simard et al., 1992]
- Correlation (kg $\sim$ pound) [Adar et al., 2009]
- **common outlinks** [Adar et al., 2009]

Cross-Language Reuse in Wikipedia



Cross-Language Reuse in Wikipedia

Currently…



Thank you!

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

References I

(2010).
Coling 2010 Organizing Committee.

Adar, E., Skinner, M., and Weld, D. (2009).
Information Arbitrage Across Multi-lingual Wikipedia.
In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM.

Barrón-Cedeño, A., Rosso, P., Agirre, E., and Labaka, G. (2010).
Plagiarism Detection across Distant Language Pairs.
In [col, 2010].

Barrón-Cedeño, A., Rosso, P., Pinto, D., and Juan, A. (2008).
On Cross-lingual Plagiarism Analysis Using a Statistical Model.
In Stein, B., Stamatatos, E., and Koppel, M., editors, *ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008)*, pages 9–13. CEUR-WS.org.

Bierce, A. (1911).
The Devil's Dictionary.
Doubleday, Page & Company.

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993).
The Mathematics of Statistical Machine Translation: Parameter Estimation.
Computational Linguistics, 19(2):263–311.

Ceska, Z., Toman, M., and Jezek, K. (2008).
Multilingual Plagiarism Detection.
In *Proceedings of the 13th International Conference on Artificial Intelligence*, pages 83–92. Springer Verlag Berlin Heidelberg.

References II

Friburger, N. and Maurel, D. (2002).
Textual similarity based on proper names.
In *SIGIR Workshop on Mathematical Formal Information Retrieval (MFIR'2002)*, pages 155–167.

HaCohen-Kerner, Y., Tayeb, A., and Ben-Dror, N. (2010).
Detection of Simple Plagiarism in Computer Science Papers.
In [col, 2010], pages 421–429.

McEnery, A. and Xiao, Z. (2007).
Parallel and Comparable Corpora: What Are They Up To?
In Rogers, M. and Anderman, G., editors, *Incorporating Corpora. The Linguist and the Translator*, pages 18–31. Clevedon.

Mcnamee, P. and Mayfield, J. (2004).
Character N-Gram Tokenization for European Language Text Retrieval.
Information Retrieval, 7(1-2):73–97.

Mohammadi, M. and GhasemAghaee, N. (2010).
Building Bilingual Parallel Corpora based on Wikipedia.
In *Second International Conference on Computer Engineering and Applications*, volume 2, pages 264–268.

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.

References III

Potthast, M., Stein, B., and Anderka, M. (2008).
A Wikipedia-Based Multilingual Retrieval Model.
In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., and White, R., editors, *30th European Conference on IR Research, ECIR 2008*, volume 4956 LNCS of *Lecture Notes in Computer Science*, pages 522–530, Berlin Heidelberg New York. Springer.

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.

Simard, M., Foster, G. F., and Isabelle, P. (1992).
Using Cognates to Align Sentences in Bilingual Corpora.
In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*.

Steinberger, R., Pouliquen, B., and Hagman, J. (2002).
Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC.
Computational Linguistics and Intelligent Text Processing. Proceedings of the CICLing 2002, 2276:415—424.

Toral, A. and Muñoz, R. (2006).
A proposal to automatically build and maintain gazetteers for Named Entity Recognition using Wikipedia.
In *Proceedings of the EACL Workshop on New Text 2006*. Association for Computational Linguistics.

Wikipedia (2010).
Party of European Socialists | Partido Socialista Europeo | Europako Alderdi Sozialista .
[Online; accessed 10-February-2010].