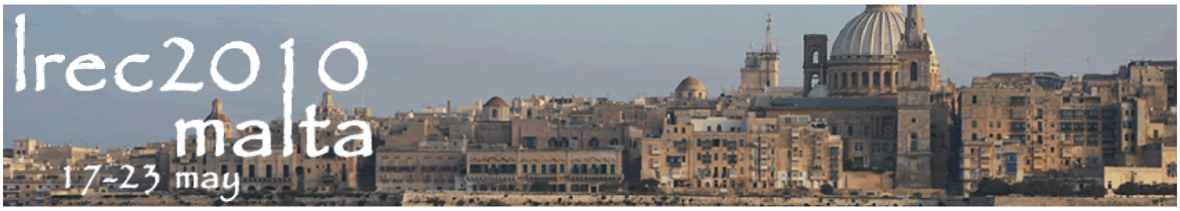
EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

# Challenges and solutions for multilingual text mining



Ralf Steinberger

http://langtech.jrc.ec.europa.eu/
http://press.jrc.it/overview.html
Ralf.Steinberger@jrc.ec.europa.eu

## What this presentation is about

- Mostly Information Extraction, but not only
- Mostly rule-based approaches, but not only

- Show the benefits of – and the need for – multilingual text processing.

- Challenge: it is an enormous effort to develop these tools.
  - For N languages, N times the effort of developing tools for one language?

- Question: Are there ways to minimise this effort?
  - Literature review
  - Our own insights

EUROPEAN COMMISSION

# Acknowledgements (alphabetical order)

## JRC colleagues (incl. former):

- Martin Atkinson
- Maud Ehrmann
- Flavio Fuart
- Erik van der Goot
- Camelia Ignat (now ECHA)
- Mijail Kabadjov
- Bruno Pouliquen (now WIPO)
- Hristo Tanev
- Vanni Zavarella
- …

## Multilingual system developers:

- Kalina Bontcheva (Sheffield University);
- Khalid Choukri (ELRA/ELDA);
- Gregory Grefenstette (Exalead);
- Frédérique Segond, Caroline Hagège and Claude Roux (Xerox Research Centre Europe);
- Aarne Ranta (Gothenburg University);
- Gregor Thurmair (Linguatec);
- Jacques Vergne (Caen University);
- Eric Wehrli (Geneva University).

Thank you !

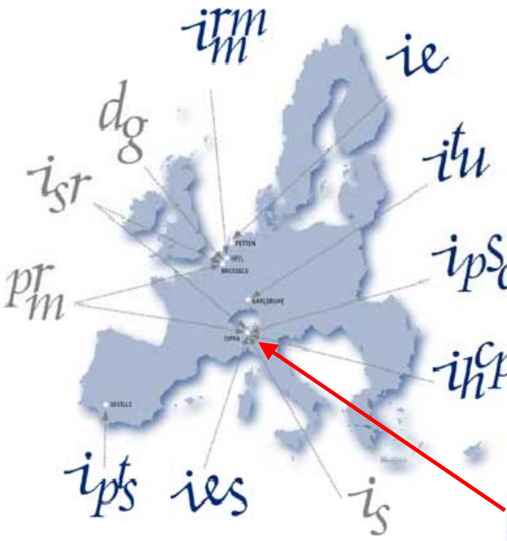
EUROPEAN COMMISSION

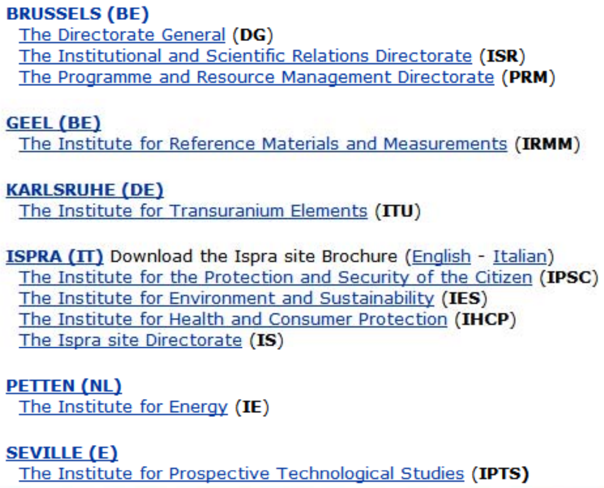
# Agenda

- European Commission – Joint Research Centre:
  - Who we are
  - Our customers and users (motivation)
- Importance of multilinguality
- Examples of multilingual EMM functionality
  - Multilingual media monitoring (publicly accessible at http://press.jrc.it/overview.html )

- How to minimise the effort of producing multilingual applications?
- Language resources that would facilitate the development of highly multilingual applications.

# Joint Research Centre - Who we are





**BRUSSELS (BE)**
The Directorate General (**DG**)
The Institutional and Scientific Relations Directorate (**ISR**)
The Programme and Resource Management Directorate (**PRM**)

**GEEL (BE)**
The Institute for Reference Materials and Measurements (**IRMM**)

**KARLSRUHE (DE)**
The Institute for Transuranium Elements (**ITU**)

**ISPRA (IT)** Download the Ispra site Brochure (English - Italian)
The Institute for the Protection and Security of the Citizen (**IPSC**)
The Institute for Environment and Sustainability (**IES**)
The Institute for Health and Consumer Protection (**IHCP**)
The Ispra site Directorate (**IS**)

**PETTEN (NL)**
The Institute for Energy (**IE**)

**SEVILLE (E)**
The Institute for Prospective Technological Studies (**IPTS**)

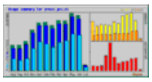
- European Commission
  (scientific-technical arm of public administration)
- Non-commercial
- Relatively small team working on Language Technology

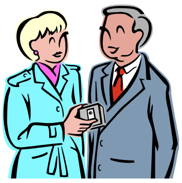
# EMM media monitoring users – wide coverage, world-wide

- European Commission (most DGs) and other EU Institutions
- EU Agencies:
  - e.g. Public Health (ECDC), Food Safety (EFSA), Chemicals Bureau (ECHA), etc.
- EU Member State organisations: e.g.
  - Public Health,
  - law enforcement authorities,
  - parliaments,
  - crisis management/humanitarian

- International and extra-European organisations: e.g.
  - various UN organisations
  - Centres for Disease Prevention and Control in the US, Canada, China, …

- The public:
  - Ca. 30,000 anonymous internet users of publicly accessible EMM systems.
  - Combined between 1 and 2 Million hits per day





JRC
EUROPEAN COMMISSION

## Agenda

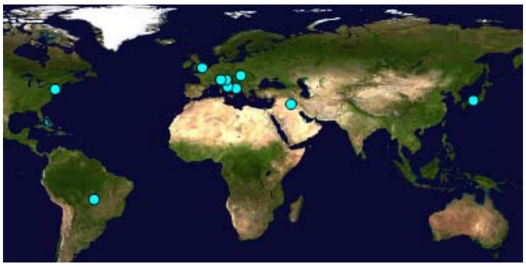
- European Commission – Joint Research Centre:
  - Who we are
  - Our customers and users (motivation)
- Importance of multilinguality
- Examples of multilingual EMM functionality
  - Multilingual media monitoring (publicly accessible at http://press.jrc.it/overview.html )

- How to minimise the effort of producing multilingual applications?
- Language resources that would facilitate the development of highly multilingual applications.

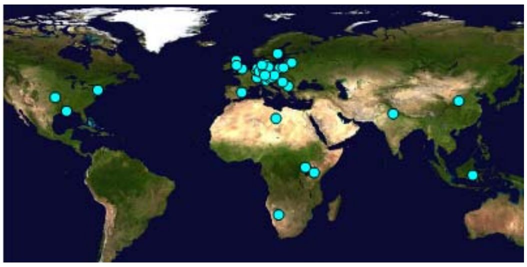
JRC
EUROPEAN COMMISSION

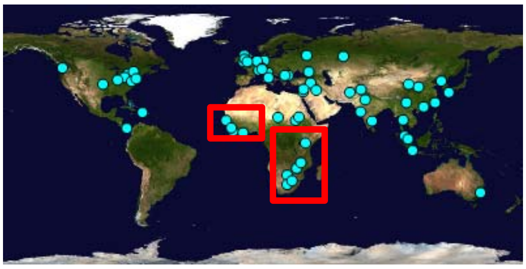
## Multilinguality: complementary news coverage in various languages

Locations mentioned in MedISys medical articles across languages – complementary coverage

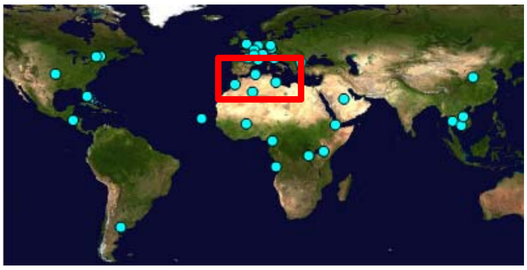


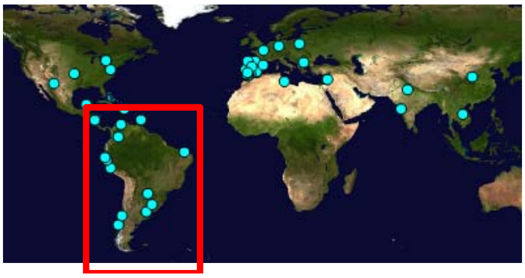
Italian - German



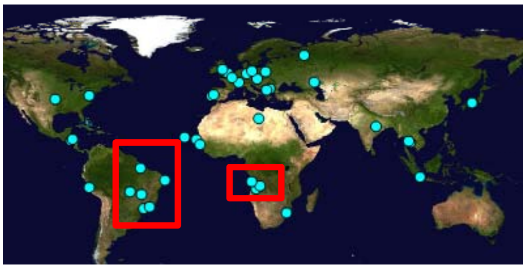


English - French

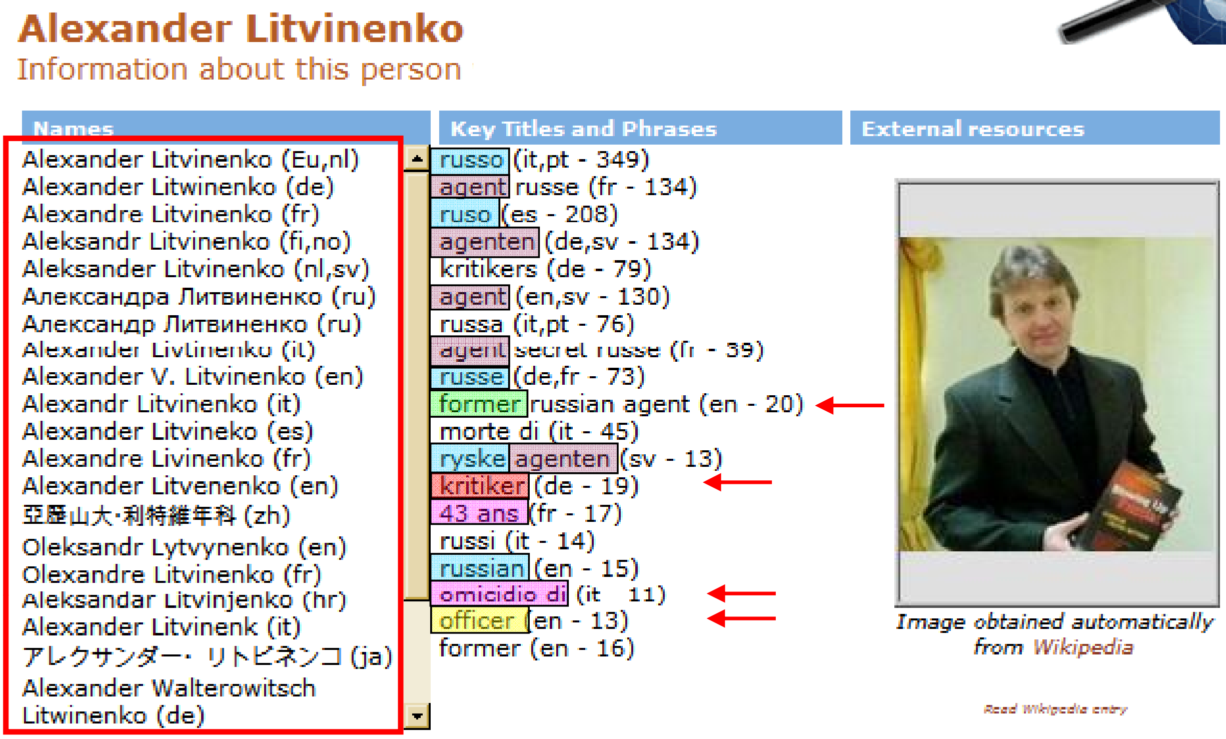




Spanish - Portuguese



# Multilinguality: Gathering *more information* about people



## Alexander Litvinenko

Information about this person

**Names**

Alexander Litvinenko (Eu,nl)
Alexander Litwinenko (de)
Alexandre Litvinenko (fr)
Aleksandr Litvinenko (fi,no)
Aleksander Litvinenko (nl,sv)
Александра Литвиненко (ru)
Александр Литвиненко (ru)
Alexander Livtinenko (it)
Alexander V. Litvinenko (en)
Alexandr Litvinenko (it)
Alexander Litvineko (es)
Alexandre Livinenko (fr)
Alexander Litvenenko (en)
亞歷山大·利特維年科 (zh)
Oleksandr Lytvynenko (en)
Olexandre Litvinenko (fr)
Aleksandar Litvinjenko (hr)
Alexander Litvinenk (it)
アレクサンダー・リトビネンコ (ja)
Alexander Walterowitsch Litwinenko (de)

**Key Titles and Phrases**

russo (it,pt - 349)
agent russe (fr - 134)
ruso (es - 208)
agenten (de,sv - 134)
kritikers (de - 79)
agent (en,sv - 130)
russa (it,pt - 76)
agent secret russe (fr - 39)
russe (de,fr - 73)
former russian agent (en - 20)
morte di (it - 45)
ryske agenten (sv - 13)
kritiker (de - 19)
43 ans (fr - 17)
russi (it - 14)
russian (en - 15)
omicidio di (it - 11)
officer (en - 13)
former (en - 16)

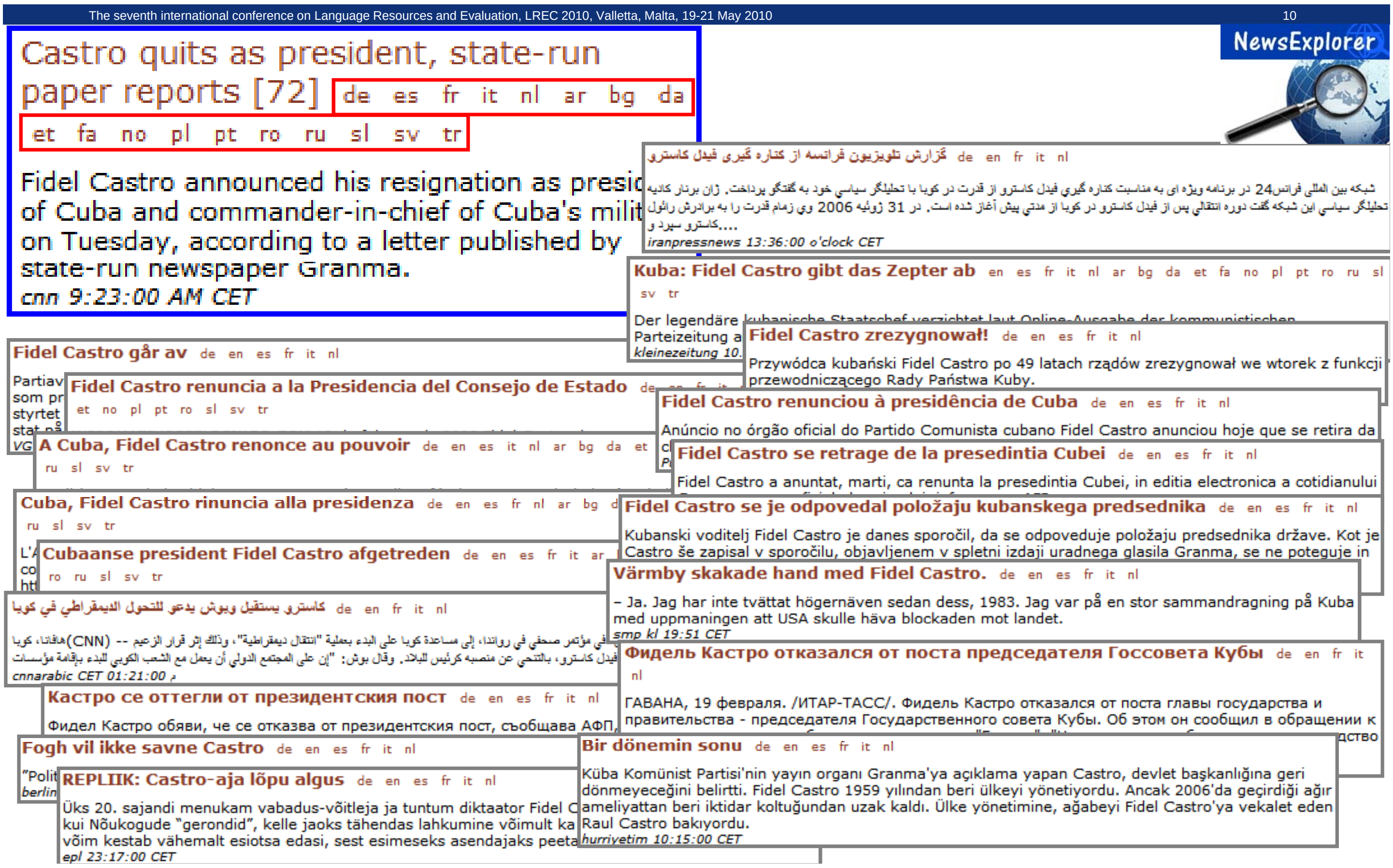
**External resources**

Image obtained automatically from Wikipedia

Read Wikipedia entry

Steinberger Ralf & Bruno Pouliquen (2009). **Cross-lingual Named Entity Recognition**. In: Satoshi Sekine & Elisabete Ranchhod (eds.): Named Entities - Recognition, Classification and Use, Benjamins Current Topics, Volume 19, pp. 137-164. John Benjamins Publishing Company.

# Multilinguality: less news bias and more transparency



Castro quits as president, state-run paper reports [72] de es fr it nl ar bg da et fa no pl pt ro ru sl sv tr

Fidel Castro announced his resignation as presid
of Cuba and commander-in-chief of Cuba's milit
on Tuesday, according to a letter published by
state-run newspaper Granma.
*cnn 9:23:00 AM CET*

**گزارش تلویزیون فرانسه از کناره گیری فیدل کاسترو** de en fr it nl

شبکه بین المللی فرانس24 در برنامه ویژه ای به مناسبت کناره گیری فیدل کاسترو از قدرت در کوبا با تحلیلگر سیاسی خود به گفتگو پرداخت. ژان برنار کادیه
تحلیلگر سیاسی این شبکه گفت دوره انتقالی پس از فیدل کاسترو در کوبا از مدتی پیش آغاز شده است. در 31 ژوئیه 2006 وی زمام قدرت را به برادرش راؤول
کاسترو سپرد و....
*iranpressnews 13:36:00 o'clock CET*

**Kuba: Fidel Castro gibt das Zepter ab** en es fr it nl ar bg da et fa no pl pt ro ru sl sv tr

Der legendäre kubanische Staatschef verzichtet laut Online-Ausgabe der kommunistischen
Parteizeitung a
*kleinezeitung 10.*

**Fidel Castro zrezygnował!** de en es fr it nl

Przywódca kubański Fidel Castro po 49 latach rządów zrezygnował we wtorek z funkcji
przewodniczącego Rady Państwa Kuby.

**Fidel Castro går av** de en es fr it nl

Partiav
som pr
styrtet
stat på
VG

**Fidel Castro renuncia a la Presidencia del Consejo de Estado** de en fr it
et no pl pt ro sl sv tr

**Fidel Castro renunciou à presidência de Cuba** de en es fr it nl

Anúncio no órgão oficial do Partido Comunista cubano Fidel Castro anunciou hoje que se retira da
c
P

**A Cuba, Fidel Castro renonce au pouvoir** de en es it nl ar bg da et
ru sl sv tr

**Fidel Castro se retrage de la presedintia Cubei** de en es fr it nl

Fidel Castro a anuntat, marti, ca renunta la presedintia Cubei, in editia electronica a cotidianului

**Cuba, Fidel Castro rinuncia alla presidenza** de en es fr nl ar bg d
ru sl sv tr

L'A
co
htt

**Fidel Castro se je odpovedal položaju kubanskega predsednika** de en es fr it nl

Kubanski voditelj Fidel Castro je danes sporočil, da se odpoveduje položaju predsednika države. Kot je
Castro še zapisal v sporočilu, objavljenem v spletni izdaji uradnega glasila Granma, se ne poteguje in

**Cubaanse president Fidel Castro afgetreden** de en es fr it ar
ro ru sl sv tr

**Värmby skakade hand med Fidel Castro.** de en es fr it nl

– Ja. Jag har inte tvättat högernäven sedan dess, 1983. Jag var på en stor sammandragning på Kuba
med uppmaningen att USA skulle häva blockaden mot landet.
*smp kl 19:51 CET*

**كاسترو يستقيل وبوش يدعو للتحول الديمقراطي في كوبا** de en fr it nl

هافانا، كوبا(CNN) -- وذلك إثر قرار الزعيم "انتقال ديمقراطية"، إلى مساعدة كوبا على البدء بعملية في مؤتمر صحفي في رواندا،
فيدل كاسترو، بالتنحي عن منصبه كرئيس للبلاد. وقال بوش: "إن على المجتمع الدولي أن يعمل مع الشعب الكوبي للبدء بإقامة مؤسسات
*cnnarabic CET 01:21:00 م*

**Фидель Кастро отказался от поста председателя Госсовета Кубы** de en fr it
nl

ГАВАНА, 19 февраля. /ИТАР-ТАСС/. Фидель Кастро отказался от поста главы государства и
правительства - председателя Государственного совета Кубы. Об этом он сообщил в обращении к
дство

**Кастро се оттегли от президентския пост** de en es fr it nl

Фидел Кастро обяви, че се отказва от президентския пост, съобщава АФП,

**Fogh vil ikke savne Castro** de en es fr it nl

"Polit
berlin

**Bir dönemin sonu** de en es fr it nl

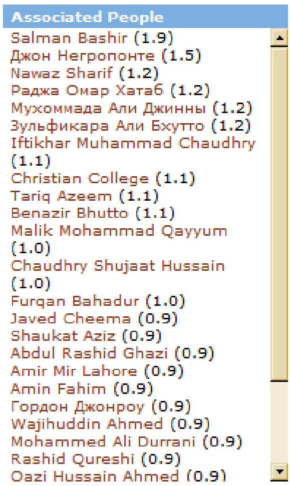
Küba Komünist Partisi'nin yayın organı Granma'ya açıklama yapan Castro, devlet başkanlığına geri
dönmeyeceğini belirtti. Fidel Castro 1959 yılından beri ülkeyi yönetiyordu. Ancak 2006'da geçirdiği ağır
ameliyattan beri iktidar koltuğundan uzak kaldı. Ülke yönetimine, ağabeyi Fidel Castro'ya vekalet eden
Raul Castro bakıyordu.
*hurriyetim 10:15:00 CET*

**REPLIIK: Castro-aja lõpu algus** de en es fr it nl

Üks 20. sajandi menukam vabadus-võitleja ja tuntum diktaator Fidel C
kui Nõukogude "gerondid", kelle jaoks tähendas lahkumine võimult ka
võim kestab vähemalt esiotsa edasi, sest esimeseks asendajaks peeta
*epl 23:17:00 CET*

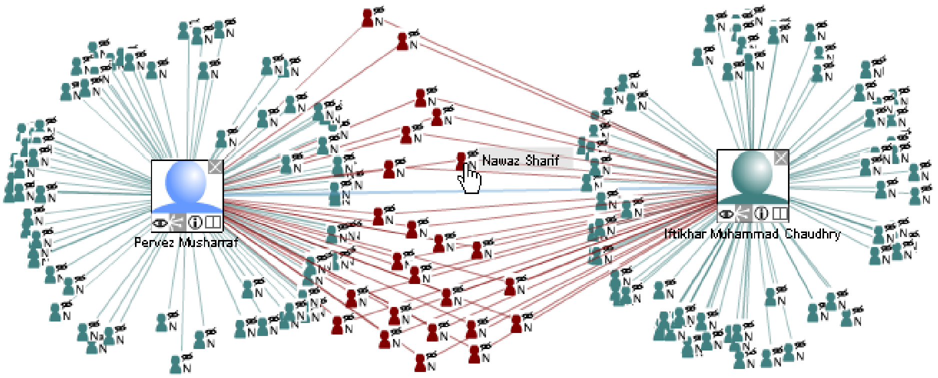
# Multilinguality: More information about relations between people

Social networks produced on the basis of many languages are **more complete** and **less biased**.



Associated People

Salman Bashir (1.9)
Джон Негропонте (1.5)
Nawaz Sharif (1.2)
Раджа Омар Хатаб (1.2)
Мухоммада Али Джинны (1.2)
Зульфикара Али Бхутто (1.2)
Iftikhar Muhammad Chaudhry (1.1)
Christian College (1.1)
Tariq Azeem (1.1)
Benazir Bhutto (1.1)
Malik Mohammad Qayyum (1.0)
Chaudhry Shujaat Hussain (1.0)
Furqan Bahadur (1.0)
Javed Cheema (0.9)
Shaukat Aziz (0.9)
Abdul Rashid Ghazi (0.9)
Amir Mir Lahore (0.9)
Amin Fahim (0.9)
Гордон Джонроу (0.9)
Wajihuddin Ahmed (0.9)
Mohammed Ali Durrani (0.9)
Rashid Qureshi (0.9)
Oazi Hussain Ahmed (0.9)



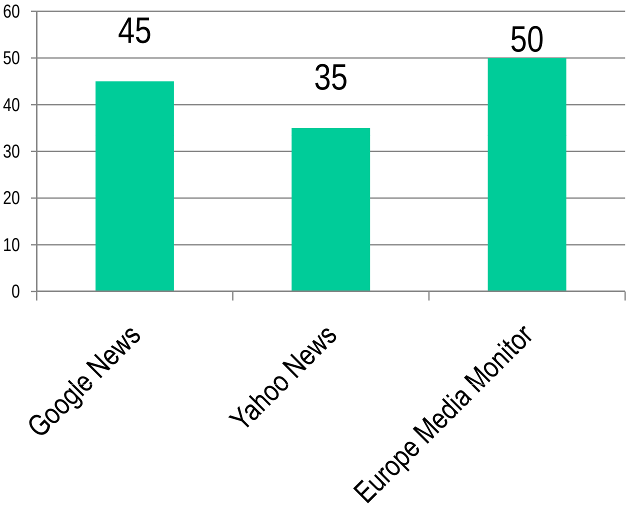
live

Pouliquen Bruno, Ralf Steinberger, Jenya Belyaeva (2007). **Multilingual multi-document continuously updated social networks.** Proceedings of the Workshop *Multi-source Multilingual Information Extraction and Summarization* (**MMIES'2007**) held at **RANLP'2007**, pp. 25-32. Borovets, Bulgaria, 26 September 2007.(**PDF**)

Hristo Tanev (2007). **Unsupervised Learning of Social Networks from a Multiple-Source News Corpus.** Proceedings of the Workshop *Multi-source Multilingual Information Extraction and Summarization* (**MMIES'2007**) held at **RANLP'2007**, pp. 33-40. Borovets, Bulgaria, 26 September 2007. (**PDF**)

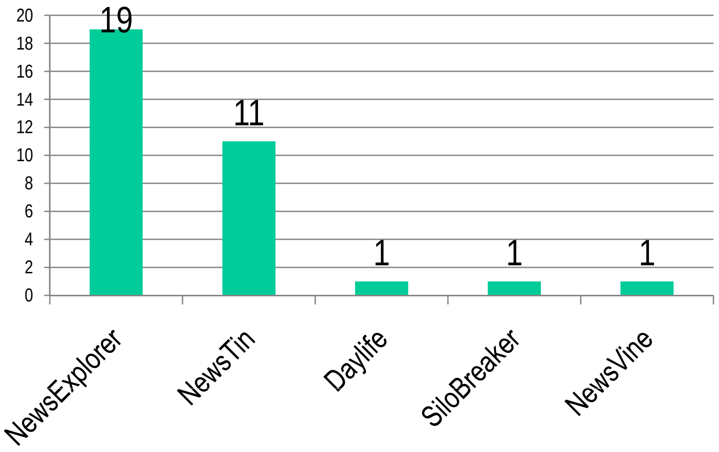
# Language coverage of various media analysis tools (March 2010)

News aggregators



in early 2008: 34, 17, 43 languages

News analysis systems



in early 2008: the same

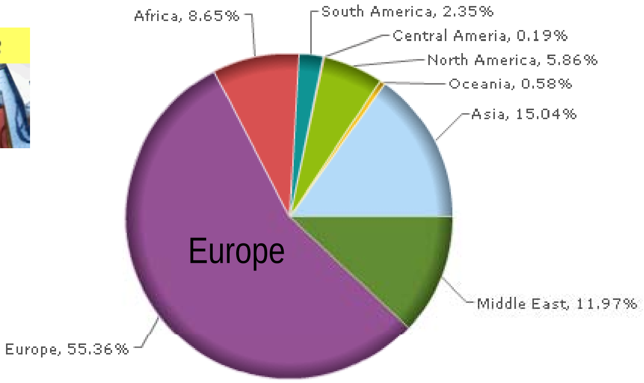
## Agenda

- European Commission – Joint Research Centre:
  - Who we are
  - Our customers and users (motivation)
- Importance of multilinguality
- Examples of multilingual EMM functionality
  - Multilingual media monitoring (publicly accessible at http://press.jrc.it/overview.html )

- How to minimise the effort of producing multilingual applications?
- Language resources that would facilitate the development of highly multilingual applications.

## What we do: 1. News gathering

- EMM news gathering engine
  - Monitors ~ 2,500 news sources
  - Gathers ~100,000 news articles per day
  - Clusters and categorises news
  - In 50 languages
  - Feeds news into the public media monitoring applications

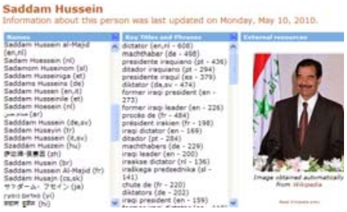


Geographical distribution of EMM news sources

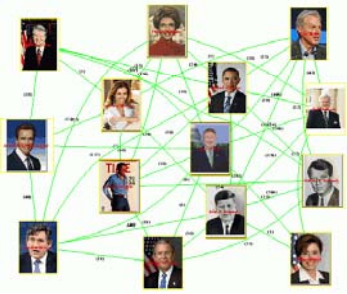
Steinberger Ralf, Bruno Pouliquen & Erik van der Goot (2009). **An Introduction to the Europe Media Monitor Family of Applications.** In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.

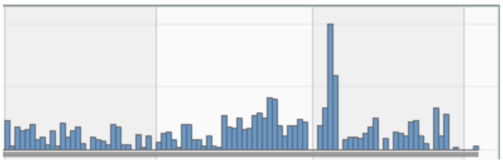
## What we do:  2. Deeper news analysis  (~20 languages)

- Breaking news detection; alerting; tracking topics over time;
- Named entity recognition and disambiguation (persons, organisations, locations);
- Name variant matching;
- Quotation recognition;
- Social network generation;
- Multi-label categorisation;
- Linking related clusters across languages;
- Event scenario template filling (6 languages);
- …


Saddam Hussein
Information about this person was last updated on Monday, May 10, 2010.

| Names | Key Titles and Phrases | External resources |
|---|---|---|
| Saddam Hussein al-Majid (en,nl)<br>Sadam Hoessein (nl)<br>Sadamom Husseinom (sl)<br>Saddam Husseiniga (et)<br>Saddams Husseins (de)<br>Saddam Hussein (en,it)<br>Saddam Husseinile (et)<br>Saddam Hoesein (nl)<br>صدام حسين (ar)<br>Sadddam Hussein (de,sv)<br>Saddam Hüseyin (tr)<br>Saddam Huassein (it,sv)<br>Szaddám Huszein (hu)<br>萨达姆·侯赛因 (zh)<br>Saddam Husain (br)<br>Saddam Hussein Al-Majid (fr)<br>Saddam Husajn (cs,sk)<br>サッダーム・フセイン (ja)<br>[illegible] (yi)<br>सद्दाम हुसैन (hi) | dictator (en,nl - 608)<br>machthaber (de - 498)<br>presidente iraquiano (pt - 436)<br>ditador iraquiano (pt - 294)<br>presidente iraquí (es - 379)<br>diktator (da,sv - 474)<br>former iraqi president (en - 273)<br>former iraqi leader (en - 226)<br>procès de (fr - 484)<br>président irakien (fr - 198)<br>iraqi dictator (en - 169)<br>ditador (pt - 284)<br>machthabers (de - 229)<br>iraqi leader (en - 200)<br>iraakse dictator (nl - 136)<br>iraškega predsednika (sl - 141)<br>chute de (fr - 220)<br>diktators (de - 202)<br>iraqi president (en - 159) | Image obtained automatically from Wikipedia |



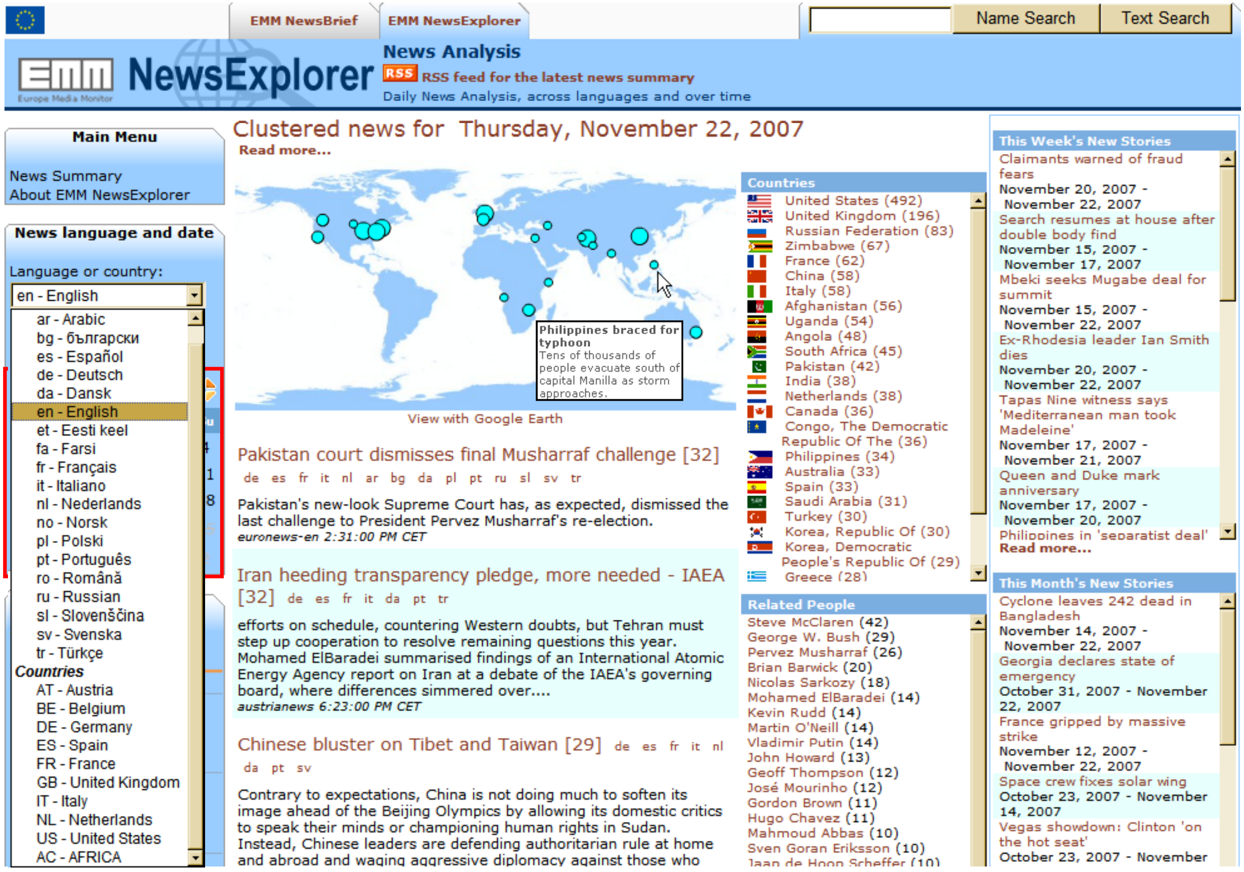


*Rice [-said]:* I would many times over liberate Iraq again from Saddam Hussein,
*CBSnews 19-MAR-10*

*Abdul Rahman [ said]:* Facts that were created after the overthrow of the Saddam Hussein regime will not be easy to maintain because there will be no US umbrella,
*guardian 04-MAR-10*

## NewsExplorer – Multilingual daily news overview

live



News Analysis
RSS feed for the latest news summary
Daily News Analysis, across languages and over time

Main Menu
News Summary
About EMM NewsExplorer

News language and date
Language or country:
en - English
ar - Arabic
bg - български
es - Español
de - Deutsch
da - Dansk
en - English
et - Eesti keel
fa - Farsi
fr - Français
it - Italiano
nl - Nederlands
no - Norsk
pl - Polski
pt - Português
ro - Română
ru - Russian
sl - Slovenščina
sv - Svenska
tr - Türkçe
*Countries*
AT - Austria
BE - Belgium
DE - Germany
ES - Spain
FR - France
GB - United Kingdom
IT - Italy
NL - Netherlands
US - United States
AC - AFRICA

Clustered news for  Thursday, November 22, 2007
**Read more...**

Philippines braced for typhoon
Tens of thousands of people evacuate south of capital Manilla as storm approaches.

View with Google Earth

Pakistan court dismisses final Musharraf challenge [32]
de es fr it nl ar bg da pl pt ru sl sv tr

Pakistan's new-look Supreme Court has, as expected, dismissed the last challenge to President Pervez Musharraf's re-election.
*euronews-en 2:31:00 PM CET*

Iran heeding transparency pledge, more needed - IAEA [32] de es fr it da pt tr

efforts on schedule, countering Western doubts, but Tehran must step up cooperation to resolve remaining questions this year. Mohamed ElBaradei summarised findings of an International Atomic Energy Agency report on Iran at a debate of the IAEA's governing board, where differences simmered over....
*austrianews 6:23:00 PM CET*

Chinese bluster on Tibet and Taiwan [29] de es fr it nl da pt sv

Contrary to expectations, China is not doing much to soften its image ahead of the Beijing Olympics by allowing its domestic critics to speak their minds or championing human rights in Sudan. Instead, Chinese leaders are defending authoritarian rule at home and abroad and waging aggressive diplomacy against those who

Countries
United States (492)
United Kingdom (196)
Russian Federation (83)
Zimbabwe (67)
France (62)
China (58)
Italy (58)
Afghanistan (56)
Uganda (54)
Angola (48)
South Africa (45)
Pakistan (42)
India (38)
Netherlands (38)
Canada (36)
Congo, The Democratic Republic Of The (36)
Philippines (34)
Australia (33)
Spain (33)
Saudi Arabia (31)
Turkey (30)
Korea, Republic Of (30)
Korea, Democratic People's Republic Of (29)
Greece (28)

Related People
Steve McClaren (42)
George W. Bush (29)
Pervez Musharraf (26)
Brian Barwick (20)
Nicolas Sarkozy (18)
Mohamed ElBaradei (14)
Kevin Rudd (14)
Martin O'Neill (14)
Vladimir Putin (14)
John Howard (13)
Geoff Thompson (12)
José Mourinho (12)
Gordon Brown (11)
Hugo Chavez (11)
Mahmoud Abbas (10)
Sven Goran Eriksson (10)
Jaap de Hoop Scheffer (10)

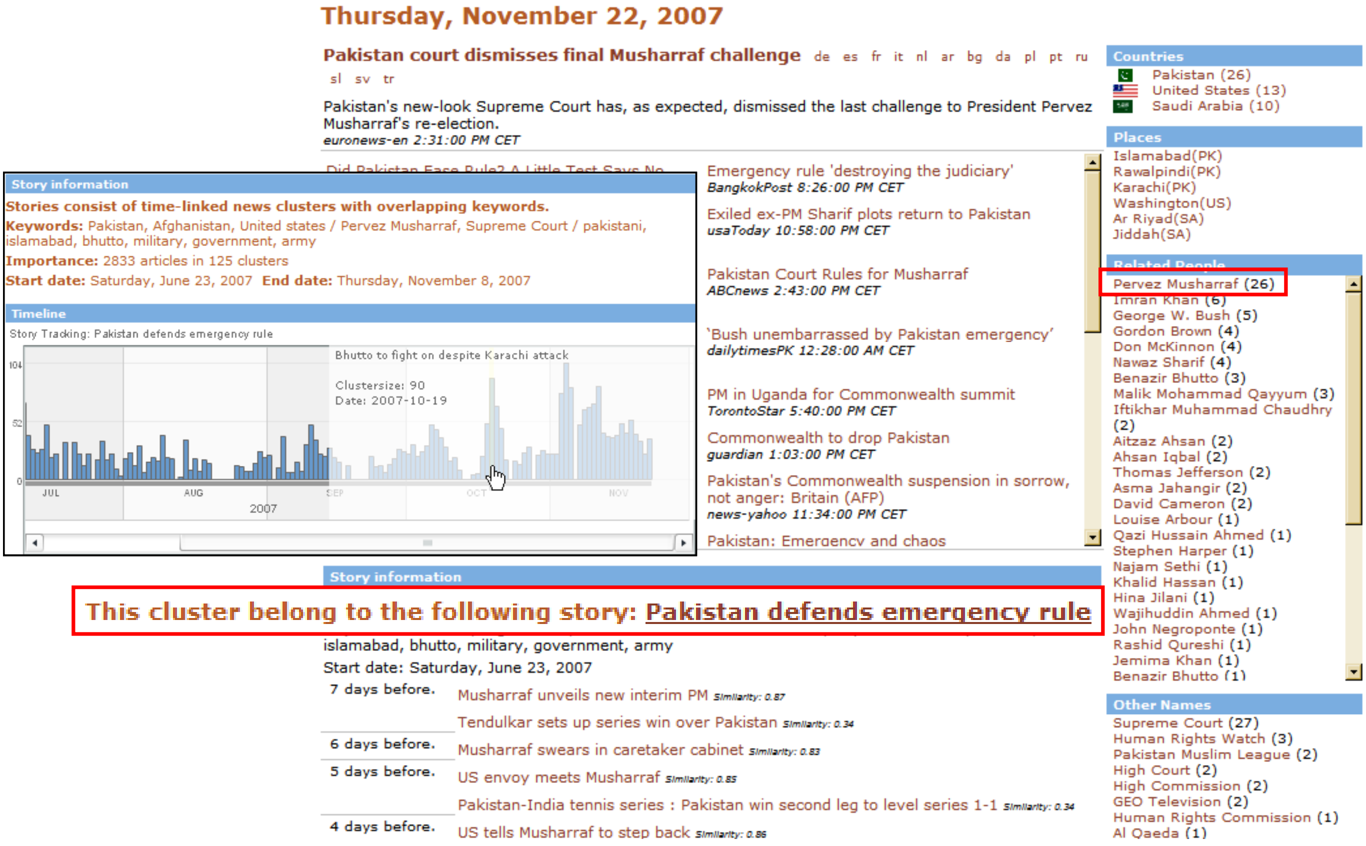
This Week's New Stories
Claimants warned of fraud fears
November 20, 2007 - November 22, 2007
Search resumes at house after double body find
November 15, 2007 - November 17, 2007
Mbeki seeks Mugabe deal for summit
November 15, 2007 - November 22, 2007
Ex-Rhodesia leader Ian Smith dies
November 20, 2007 - November 22, 2007
Tapas Nine witness says 'Mediterranean man took Madeleine'
November 17, 2007 - November 21, 2007
Queen and Duke mark anniversary
November 17, 2007 - November 20, 2007
Philippines in 'separatist deal'
**Read more...**

This Month's New Stories
Cyclone leaves 242 dead in Bangladesh
November 14, 2007 - November 22, 2007
Georgia declares state of emergency
October 31, 2007 - November 22, 2007
France gripped by massive strike
November 12, 2007 - November 22, 2007
Space crew fixes solar wing
October 23, 2007 - November 14, 2007
Vegas showdown: Clinton 'on the hot seat'
October 23, 2007 - November

JRC
EUROPEAN COMMISSION

# NewsExplorer – Aggregation of clusters into longer ‘stories’

live



Thursday, November 22, 2007

Pakistan court dismisses final Musharraf challenge de es fr it nl ar bg da pl pt ru sl sv tr

Pakistan's new-look Supreme Court has, as expected, dismissed the last challenge to President Pervez Musharraf's re-election.
*euronews-en 2:31:00 PM CET*

Story information

**Stories consist of time-linked news clusters with overlapping keywords.**

**Keywords:** Pakistan, Afghanistan, United states / Pervez Musharraf, Supreme Court / pakistani, islamabad, bhutto, military, government, army

**Importance:** 2833 articles in 125 clusters

**Start date:** Saturday, June 23, 2007 **End date:** Thursday, November 8, 2007

Timeline

Story Tracking: Pakistan defends emergency rule
Bhutto to fight on despite Karachi attack
Clustersize: 90
Date: 2007-10-19
104
52
0
JUL
AUG
SEP
OCT
NOV
2007

Emergency rule 'destroying the judiciary'
*BangkokPost 8:26:00 PM CET*

Exiled ex-PM Sharif plots return to Pakistan
*usaToday 10:58:00 PM CET*

Pakistan Court Rules for Musharraf
*ABCnews 2:43:00 PM CET*

‘Bush unembarrassed by Pakistan emergency’
*dailytimesPK 12:28:00 AM CET*

PM in Uganda for Commonwealth summit
*TorontoStar 5:40:00 PM CET*

Commonwealth to drop Pakistan
*guardian 1:03:00 PM CET*

Pakistan's Commonwealth suspension in sorrow, not anger: Britain (AFP)
*news-yahoo 11:34:00 PM CET*

Pakistan: Emergency and chaos

Story information

**This cluster belong to the following story: Pakistan defends emergency rule**

islamabad, bhutto, military, government, army

Start date: Saturday, June 23, 2007

| | |
|---|---|
| 7 days before. | Musharraf unveils new interim PM *Similarity: 0.87* |
| | Tendulkar sets up series win over Pakistan *Similarity: 0.34* |
| 6 days before. | Musharraf swears in caretaker cabinet *Similarity: 0.83* |
| 5 days before. | US envoy meets Musharraf *Similarity: 0.85* |
| | Pakistan-India tennis series : Pakistan win second leg to level series 1-1 *Similarity: 0.34* |
| 4 days before. | US tells Musharraf to step back *Similarity: 0.86* |

Countries

| | |
|---|---|
| Pakistan | (26) |
| United States | (13) |
| Saudi Arabia | (10) |

Places

Islamabad(PK)
Rawalpindi(PK)
Karachi(PK)
Washington(US)
Ar Riyad(SA)
Jiddah(SA)

Related People

| | |
|---|---|
| Pervez Musharraf | (26) |
| Imran Khan | (6) |
| George W. Bush | (5) |
| Gordon Brown | (4) |
| Don McKinnon | (4) |
| Nawaz Sharif | (4) |
| Benazir Bhutto | (3) |
| Malik Muhammad Qayyum | (3) |
| Iftikhar Muhammad Chaudhry | (2) |
| Aitzaz Ahsan | (2) |
| Ahsan Iqbal | (2) |
| Thomas Jefferson | (2) |
| Asma Jahangir | (2) |
| David Cameron | (2) |
| Louise Arbour | (1) |
| Qazi Hussain Ahmed | (1) |
| Stephen Harper | (1) |
| Najam Sethi | (1) |
| Khalid Hassan | (1) |
| Hina Jilani | (1) |
| Wajihuddin Ahmed | (1) |
| John Negroponte | (1) |
| Rashid Qureshi | (1) |
| Jemima Khan | (1) |
| Benazir Bhutto | (1) |

Other Names

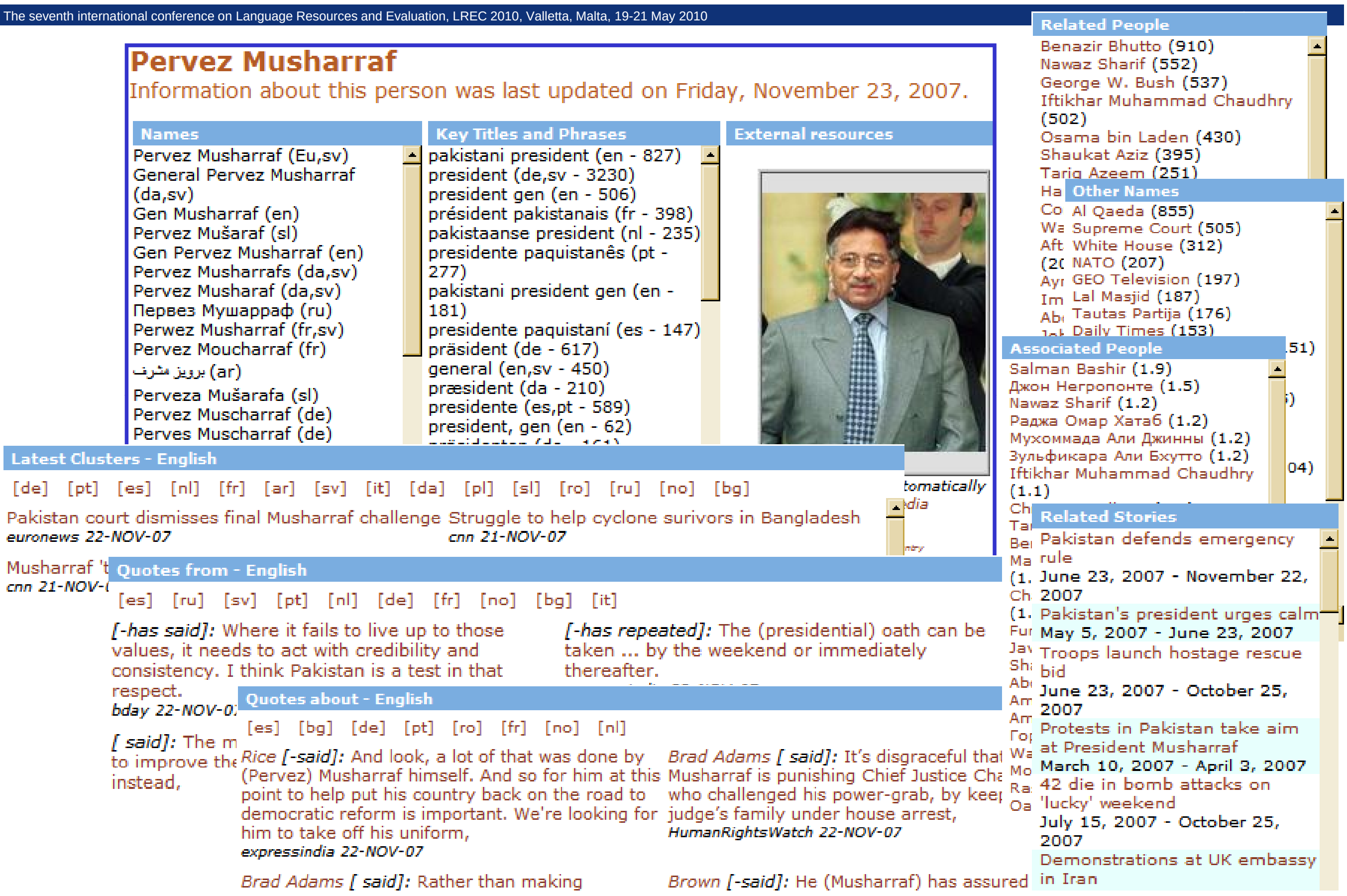
| | |
|---|---|
| Supreme Court | (27) |
| Human Rights Watch | (3) |
| Pakistan Muslim League | (2) |
| High Court | (2) |
| High Commission | (2) |
| GEO Television | (2) |
| Human Rights Commission | (1) |
| Al Qaeda | (1) |

JRC
EUROPEAN COMMISSION

# NewsExplorer –Information about people

## collected from multiple languages and over time

live



Pervez Musharraf

Information about this person was last updated on Friday, November 23, 2007.

Names

Pervez Musharraf (Eu,sv)
General Pervez Musharraf (da,sv)
Gen Musharraf (en)
Pervez Mušaraf (sl)
Gen Pervez Musharraf (en)
Pervez Musharrafs (da,sv)
Pervez Musharaf (da,sv)
Первез Мушаррафа (ru)
Perwez Musharraf (fr,sv)
Pervez Moucharraf (fr)
برويز مشرف (ar)
Perveza Mušarafa (sl)
Pervez Muscharraf (de)
Perves Muscharraf (de)

Key Titles and Phrases

| | |
|---|---|
| pakistani president | (en - 827) |
| president | (de,sv - 3230) |
| president gen | (en - 506) |
| président pakistanais | (fr - 398) |
| pakistaanse president | (nl - 235) |
| presidente paquistanês | (pt - 277) |
| pakistani president gen | (en - 181) |
| presidente paquistaní | (es - 147) |
| präsident | (de - 617) |
| general | (en,sv - 450) |
| præsident | (da - 210) |
| presidente | (es,pt - 589) |
| president, gen | (en - 62) |

External resources

Related People

| | |
|---|---|
| Benazir Bhutto | (910) |
| Nawaz Sharif | (552) |
| George W. Bush | (537) |
| Iftikhar Muhammad Chaudhry | (502) |
| Osama bin Laden | (430) |
| Shaukat Aziz | (395) |
| Tariq Azeem | (251) |

Other Names

| | |
|---|---|
| Al Qaeda | (855) |
| Supreme Court | (505) |
| White House | (312) |
| NATO | (207) |
| GEO Television | (197) |
| Lal Masjid | (187) |
| Tautas Partija | (176) |
| Daily Times | (153) |

Associated People

| | |
|---|---|
| Salman Bashir | (1.9) |
| Джон Негропонте | (1.5) |
| Nawaz Sharif | (1.2) |
| Раджа Омар Хатаб | (1.2) |
| Мухоммада Али Джинны | (1.2) |
| Зульфикара Али Бхутто | (1.2) |
| Iftikhar Muhammad Chaudhry | (1.1) |

Related Stories

| | |
|---|---|
| Pakistan defends emergency rule | June 23, 2007 - November 22, 2007 |
| Pakistan's president urges calm | May 5, 2007 - June 23, 2007 |
| Troops launch hostage rescue bid | June 23, 2007 - October 25, 2007 |
| Protests in Pakistan take aim at President Musharraf | March 10, 2007 - April 3, 2007 |
| 42 die in bomb attacks on 'lucky' weekend | July 15, 2007 - October 25, 2007 |
| Demonstrations at UK embassy in Iran | |

Latest Clusters - English

[de] [pt] [es] [nl] [fr] [ar] [sv] [it] [da] [pl] [sl] [ro] [ru] [no] [bg]

Pakistan court dismisses final Musharraf challenge
*euronews 22-NOV-07*

Struggle to help cyclone surivors in Bangladesh
*cnn 21-NOV-07*

Quotes from - English

[es] [ru] [sv] [pt] [nl] [de] [fr] [no] [bg] [it]

*[-has said]:* Where it fails to live up to those values, it needs to act with credibility and consistency. I think Pakistan is a test in that respect.

*[-has repeated]:* The (presidential) oath can be taken ... by the weekend or immediately thereafter.

Quotes about - English

[es] [bg] [de] [pt] [ro] [fr] [no] [nl]

*Rice [-said]:* And look, a lot of that was done by (Pervez) Musharraf himself. And so for him at this point to help put his country back on the road to democratic reform is important. We're looking for him to take off his uniform,
*expressindia 22-NOV-07*

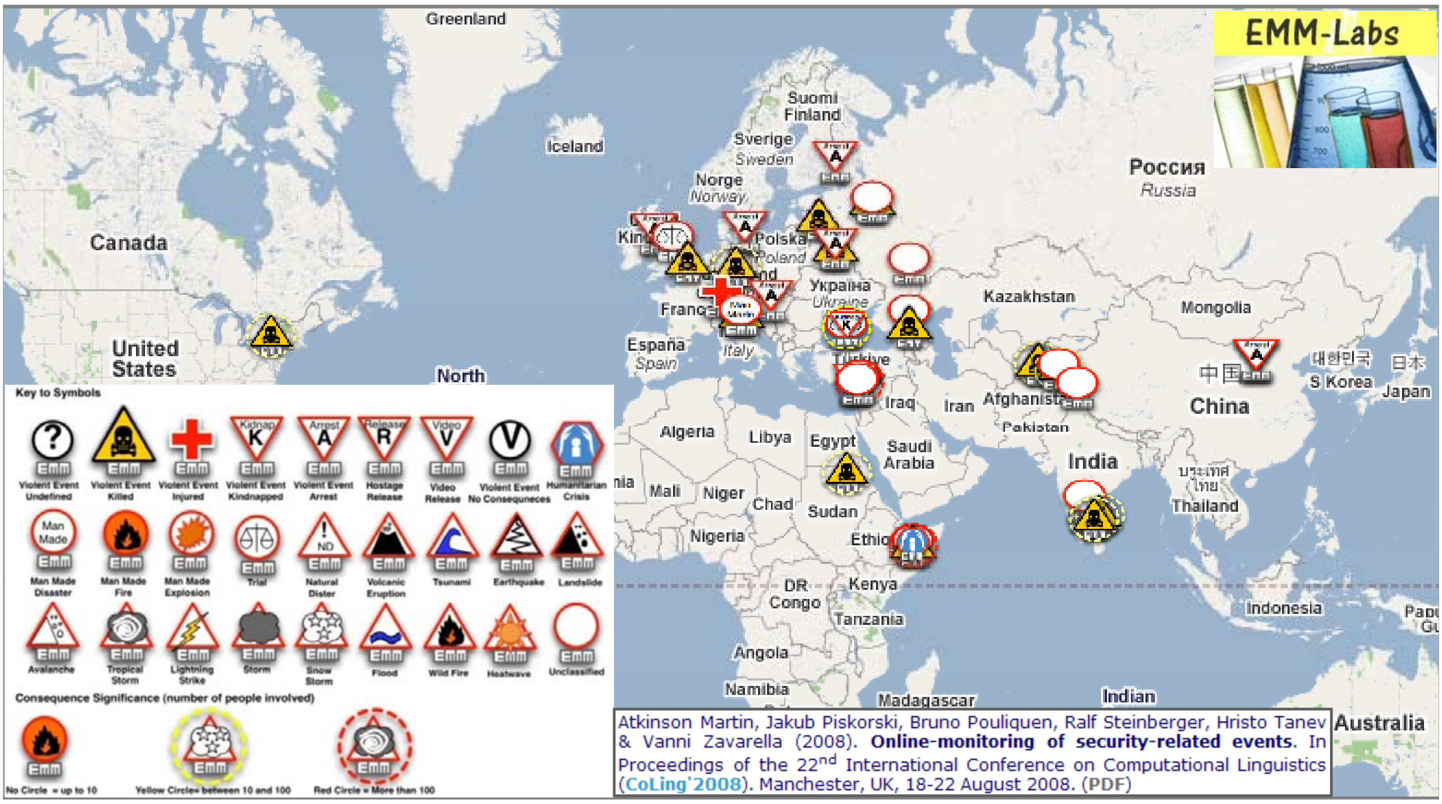
*Brad Adams [ said]:* It’s disgraceful that Musharraf is punishing Chief Justice Cha who challenged his power-grab, by keep judge’s family under house arrest,
*HumanRightsWatch 22-NOV-07*

*Brad Adams [ said]:* Rather than making

*Brown [-said]:* He (Musharraf) has assured

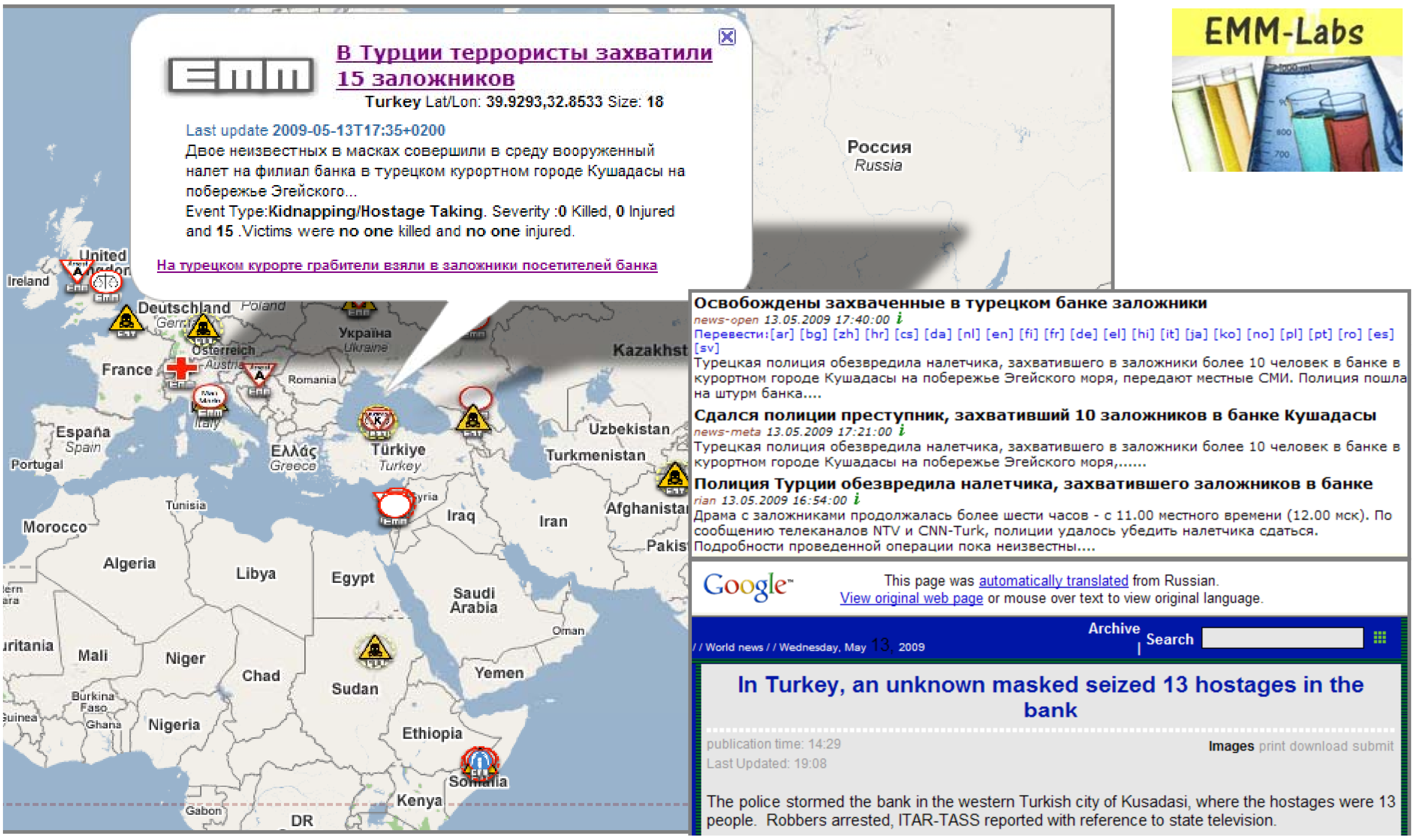
# Detection and visualisation of events (violence/disasters/humanitarian/…)

**Objective**: global crisis monitoring. **Languages**: En, Fr, Es, It, Pt, Ru + (Ar)



Atkinson Martin, Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger, Hristo Tanev & Vanni Zavarella (2008). **Online-monitoring of security-related events.** In Proceedings of the 22nd International Conference on Computational Linguistics (CoLing'2008). Manchester, UK, 18-22 August 2008. (PDF)

# Detection and visualisation of events (violence/disasters/humanitarian/…)



**В Турции террористы захватили 15 заложников**

**Turkey** Lat/Lon: **39.9293,32.8533** Size: **18**

Last update **2009-05-13T17:35+0200**

Двое неизвестных в масках совершили в среду вооруженный налет на филиал банка в турецком курортном городе Кушадасы на побережье Эгейского...

Event Type:**Kidnapping/Hostage Taking**. Severity :**0** Killed, **0** Injured and **15** .Victims were **no one** killed and **no one** injured.

На турецком курорте грабители взяли в заложники посетителей банка

**Освобождены захваченные в турецком банке заложники**

*news-open 13.05.2009 17:40:00*

Перевести:[ar] [bg] [zh] [hr] [cs] [da] [nl] [en] [fi] [fr] [de] [el] [hi] [it] [ja] [ko] [no] [pl] [pt] [ro] [es] [sv]

Турецкая полиция обезвредила налетчика, захватившего в заложники более 10 человек в банке в курортном городе Кушадасы на побережье Эгейского моря, передают местные СМИ. Полиция пошла на штурм банка....

**Сдался полиции преступник, захвативший 10 заложников в банке Кушадасы**

*news-meta 13.05.2009 17:21:00*

Турецкая полиция обезвредила налетчика, захватившего в заложники более 10 человек в банке в курортном городе Кушадасы на побережье Эгейского моря,......

**Полиция Турции обезвредила налетчика, захватившего заложников в банке**

*rian 13.05.2009 16:54:00*

Драма с заложниками продолжалась более шести часов - с 11.00 местного времени (12.00 мск). По сообщению телеканалов NTV и CNN-Turk, полиции удалось убедить налетчика сдаться. Подробности проведенной операции пока неизвестны....

This page was automatically translated from Russian. View original web page or mouse over text to view original language.

/ / World news / / Wednesday, May 13, 2009 Archive | Search

**In Turkey, an unknown masked seized 13 hostages in the bank**

publication time: 14:29
Last Updated: 19:08

Images print download submit

The police stormed the bank in the western Turkish city of Kusadasi, where the hostages were 13 people. Robbers arrested, ITAR-TASS reported with reference to state television.

# Agenda

- European Commission – Joint Research Centre:
  - Who we are
  - Our customers and users (motivation)
- Importance of multilinguality
- Examples of multilingual EMM functionality
  - Multilingual media monitoring (publicly accessible at http://press.jrc.it/overview.html )

- How to minimise the effort of producing multilingual applications?
- Language resources that would facilitate the development of highly multilingual applications.

# Insights collected from various teams

Effort for N languages = N times effort for one language? How to save effort?

- Depends on required level of analysis
- Complex applications may require deeper linguistic analysis
- But: even simple means can take you relatively far
- Recognition easier than generation (e.g. agreement)

- Typical and most common approach for rule-based systems:
  - Develop in one language
  - Reuse resources and adapt to new languages
  - E.g. Gamon et al. (1997); Rayner & Bouillon (1996), Pastra et el. (2002); Carenini et al. (2007); Maynard et al. (2003)

# Insights collected from various teams; Guidelines; Ideas (1)

1. Use Unicode (Maynard et al. 2002)

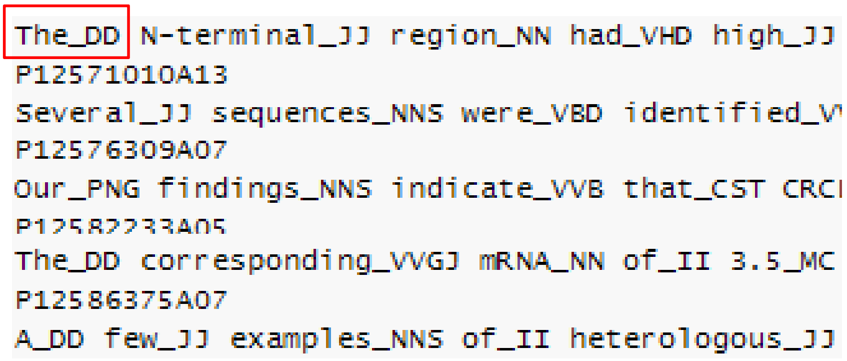
2. Use virtual keyboards to enter foreign language data (Maynard et al. 2002)

3. Modularity (Pastra et al. 2002; Maynard et al. 2002)

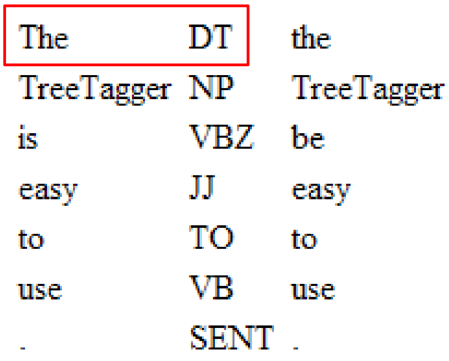
# Insights collected from various teams (2)

4. Shared token classes, ideally based on surface features (Bering et al. 2003)
   - E.g. case information, includes-number, includes hyphen; string length

5. Uniform input and output structures (Carenini et al. 2007; Bering et al. 2003)
   - Data format
   - When possible: same part-of-speech classes, same grammatical categories



```
The_DD N-terminal_JJ region_NN had_VHD high_JJ
P12571010A13
Several_JJ sequences_NNS were_VBD identified_V
P12576309A07
Our_PNG findings_NNS indicate_VVB that_CST CRC
P12582233A05
The_DD corresponding_VVGJ mRNA_NN of_II 3.5_MC
P12586375A07
A_DD few_JJ examples_NNS of_II heterologous_JJ
```



```
The         DT     the
TreeTagger  NP     TreeTagger
is          VBZ    be
easy        JJ     easy
to          TO     to
use         VB     use
.           SENT   .
```

JRC
EUROPEAN COMMISSION

## Insights collected from various teams (3)

6. Simplicity of rules and the lexicon (Carenini et al. 2007; Vergne 2002)

- E.g. Identification of subject-verb pairs for 5 languages (Vergne 2002)
  - <200 dictionary elements per language
  - Case information
  - Regular expressions matching certain combinations of word endings

- E.g. Chunker for 23 languages (Vergne 2009), using only
  - String length information
  - Word frequency

- E.g. under-specification (agreement; order of modifiers) in recognition of quotation and speakers (Pouliquen et al. 2007)
  - "…" said the former 56-year-old British Prime Minister *Tony Blair*

JRC
EUROPEAN COMMISSION

## Insights collected from various teams (4)

7. Share resources between languages (lexica, gazetteers, grammar rules) (Bering et al. 2003)

- Language-independent rule set + language-specific rules, e.g. for date recognition
  - 20.10.2010 (generic)
  - 20th of October of the year 2010 (language-specific)
    (Ignat et al. 2003 also cover this and more cases with generic rules)

8. Use theory-neutral date types (Pastra et al. 2002; Maynard et al. 2002)

- For the Language Engineering architecture GATE

# Insights collected from various teams (5)

9. Adhere to a grammar theory, e.g.
   - Bender & Flickinger (2005) use **HPSG** for general-purpose grammars;
   - Gamon et al. (1997) adhere to **Universal Grammar** for generic Microsoft NLP grammar;
   - Wehrli (2007) uses **Chomsky's generative grammar** to build parsers;
   - Ranta (2009, e.g. pp. 47ff; LREC tutorial) works within the **Grammatical Framework** GF.

Benefits mentioned:

- The mere existence of an abstract syntax implies grammar sharing;
- Creating a generic grammar and parameterise it to handle many languages;
- Shared grammars by language group;
- Treating some linguistic phenomena in a systematic way
  (e.g. clitics; morphological agreement; phrase ordering; …);
- Generating starter grammars based on linguistic features (Bender & Flickinger 2005).

# The promising contribution of Machine Learning (ML)

- Idea: self-learning software learns rules and vocabulary, e.g. from examples

- Success story: **Statistical Machine Translation** (SMT);
  - Google: 57 languages (incl. 6 alpha versions); 1596 language pairs (May 2010).

- ML for **Named Entity Recognition** (NER) (Nadeau & Sekine 2009)
  - Supervised ML: train on previously annotated corpora
    - Challenge: Corpora annotation is labour-intensive and expensive
  - Semi-supervised, weakly-supervised ML:
    - Use set of human-provided seeds to start learning process
    - Use boot-strapping to increase number of patterns and resources
  - Unsupervised ML:
    - E.g. words appearing synchronously in news articles (Shinyama & Sekine 2004).

# Machine Learning versus manual intervention

- Issue: how to combine ML methods with manual intervention, e.g. to correct
  - E.g. Output of SVM or HMM is difficult to modify manually.

→ EMM approach:

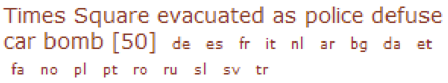
  - Use hand-crafted rules;
  - Use external knowledge sources (dictionaries) when available;
  - Use bootstrapping and ML methods to enhance these dictionaries;
  - Empirical testing.

- Benefit:
  - Keep control over the recognition performance
  - Spend less time per language than when using ML: 1 week – 3 months per language to
    - add news sources,
    - translate Boolean category definitions
    - Add linguistic IE resources (to recognise persons, organisations, locations, quotations, dates)
    - Evaluation

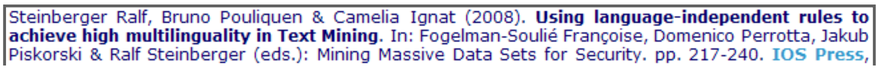
# EMM guidelines

- Keep applications simple!
  - Under-specify (constraints are time-consuming to produce and may hinder you in other languages)
  - Use bags of words without specifying agreement and order, if possible
  - Don't disambiguate if you can avoid it
  - No grammar theory

- Use language-independent rules, if possible
- Use as little language-specific resources as possible (POS taggers, parsers, dictionaries, …)
- Modularity: keep language-specific resources outside the rules → plug in any new language
- Do not use language pair-specific resources
  - NewsExplorer covers 20 languages, 190 language pairs



Times Square evacuated as police defuse car bomb [50] de es fr it nl ar bg da et fa no pl pt ro ru sl sv tr



Steinberger Ralf, Bruno Pouliquen & Camelia Ignat (2008). **Using language-independent rules to achieve high multilinguality in Text Mining**. In: Fogelman-Soulié Françoise, Domenico Perrotta, Jakub Piskorski & Ralf Steinberger (eds.): Mining Massive Data Sets for Security. pp. 217-240. IOS Press,

## Concretely, …

- What we (JRC) cannot use:
  - Syntactic parsers
  - Part-of-speech taggers
  - Full dictionary for any of the languages
- What we do use:
  - Targeted world lists
    - Name titles, etc.
    - Gazetteers of place names and locations
    - Sentiment words
    - Reporting verbs
    - Stop word lists!!!
  - Light-weight suffix stripping rules and generation of morphological variants
  - Boolean combinations of category-defining words (for document classification)
  - The output of our own NER tools
  - Machine Learning and boot-strapping

## Agenda

- European Commission – Joint Research Centre:
  - Who we are
  - Our customers and users (motivation)
- Importance of multilinguality
- Examples of multilingual EMM functionality
  - Multilingual media monitoring (publicly accessible at http://press.jrc.it/overview.html )

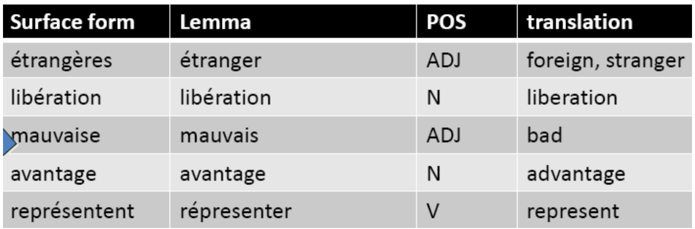
- How to minimise the effort of producing multilingual applications?
- Language resources that would facilitate the development of highly multilingual applications.

# EMM insight: Need for multilingual resources

→ Need for uniform highly multilingual resources for information extraction, etc.

Similar to Multext and Multext-East, EuroWordNet, …

- *Parallel* lexical resources, e.g.
  - Multilingual **gazetteers** for geo-coding (e.g. GeoNames)
  - Multilingual **dictionaries** using the same features and format (MulText, Euro-WordNet)
- Multilingual **taggers and parsers** producing the same output type and output format.
- Annotated **multilingual parallel text corpora** (both for training and for testing) (e.g. JRC-Acquis)
- **Single access point for licensing** issues (e.g. ELRA / LDC)
- Ideally **freely available** (see also: G. Grefenstette's recommendation at recent FLaReNet workshop: http://www.flarenet.eu/?q=node/347)



| Surface form | Lemma | POS | translation |
|---|---|---|---|
| étrangères | étranger | ADJ | foreign, stranger |
| libération | libération | N | liberation |
| mauvaise | mauvais | ADJ | bad |
| avantage | avantage | N | advantage |
| représentent | répresenter | V | represent |

# Summary

# Summary

- Benefits of multilinguality: capture complementary coverage across languages
  - Contents
  - Opinions
- Major challenge: time needed to develop resources and applications
- Proposals how to keep the effort down:
  - Unicode
  - Virtual keyboards
  - Modularity
  - Shared token classes (surface features)
  - Uniform input and output structures
  - Simplicity of rules and the lexicon
  - Share resources between languages
  - Theory-neutral / adhere to grammar theory
  - Use Machine Learning
  - Under-specification
  - Minimise use of language-specific resources
  - Avoid language pair-specific resources

# Summary (2)

- Useful language resources to achieve high multilinguality
  - *Uniform and parallel* dictionaries, corpora and tools.
- Surely, there are more unpublished insights → please share them with me.

Ralf.Steinberger@jrc.ec.europa.eu

- Different applications require different means and there are many ways of doing things.