



Linking News Content Across Languages

NODALIDA 2009

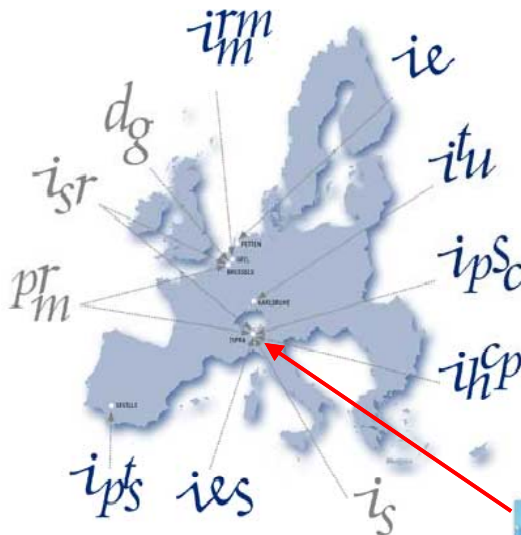
Odense, Denmark, 16 May 2009

Ralf Steinberger

& the JRC's *OPTIMA* team
(Open Source Text Information Mining and Analysis)

<http://langtech.jrc.it/>
<http://press.jrc.it/overview.html>

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 \(IPTS\)](#)



- Europe Media Monitor (EMM) applications
 - Publicly accessible at <http://press.jrc.it/overview.html>

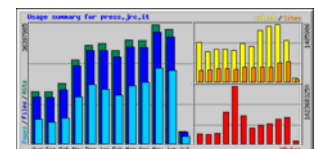


- NewsExplorer functionality
- NewsExplorer language technology components
 - Multilingual person name recognition
 - Name variant matching
 - Cross-lingual linking of news clusters
- Summary and Ongoing work

- EMM news gathering engine
 - Monitors ~ 2,200 news sources
 - Gathers 80,000 – 100,000 news articles per day
 - In about 43 languages
 - Visits some sites every 10 minutes
 - Extracts text from the web page
 - Converts text into Unicode-encoded RSS
 - Feeds the news into the four publicly accessible media monitoring systems



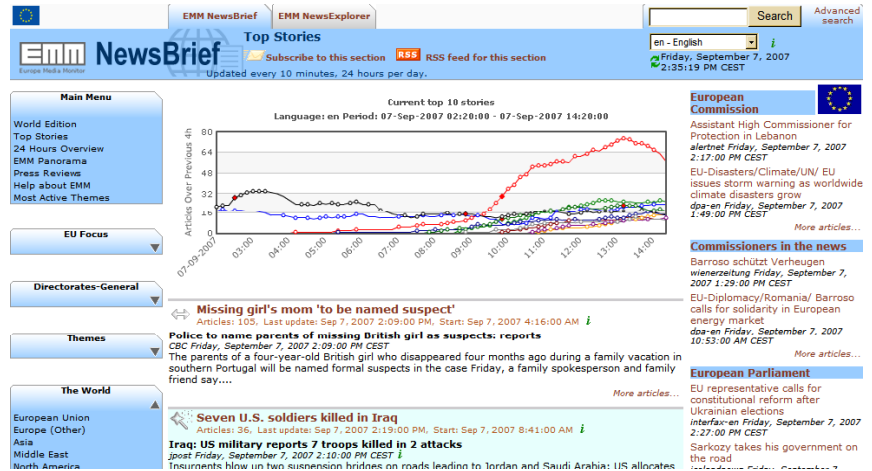
- Combined between 1 and 2 Million hits per day
- 30,000 – 50,000 distinct users per day





- Public site: <http://press.jrc.it/NewsBrief/>
- Categorises news into ~ 600 categories, using:
 - Boolean search word combinations
 - vicinity operators
 - optional weights
 - regular expressions

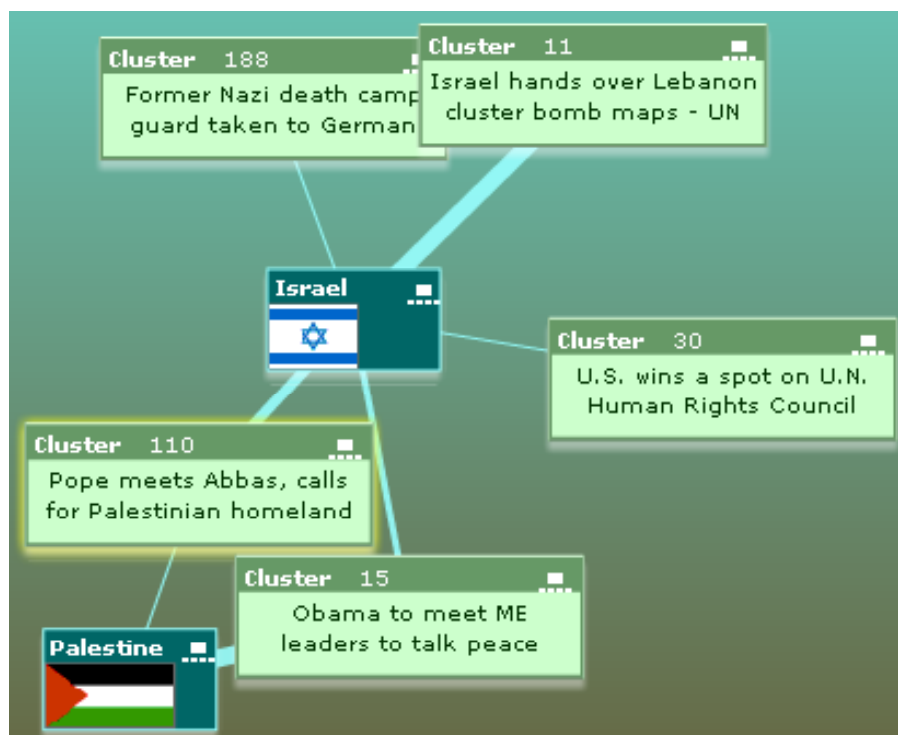
- Clusters and tracks news live (multi-lingually)
- Sends out email notifications for each category
- Detects breaking news
- Short-term story tracking



Atkinson Martin & Erik Van der Goot (2009). **Near Real Time Information Mining in Multilingual News**. Proceedings of the 18th International World Wide Web Conference (WWW'2009), pp. 1153-1154. Madrid, 20-24 April 2009. (PDF)

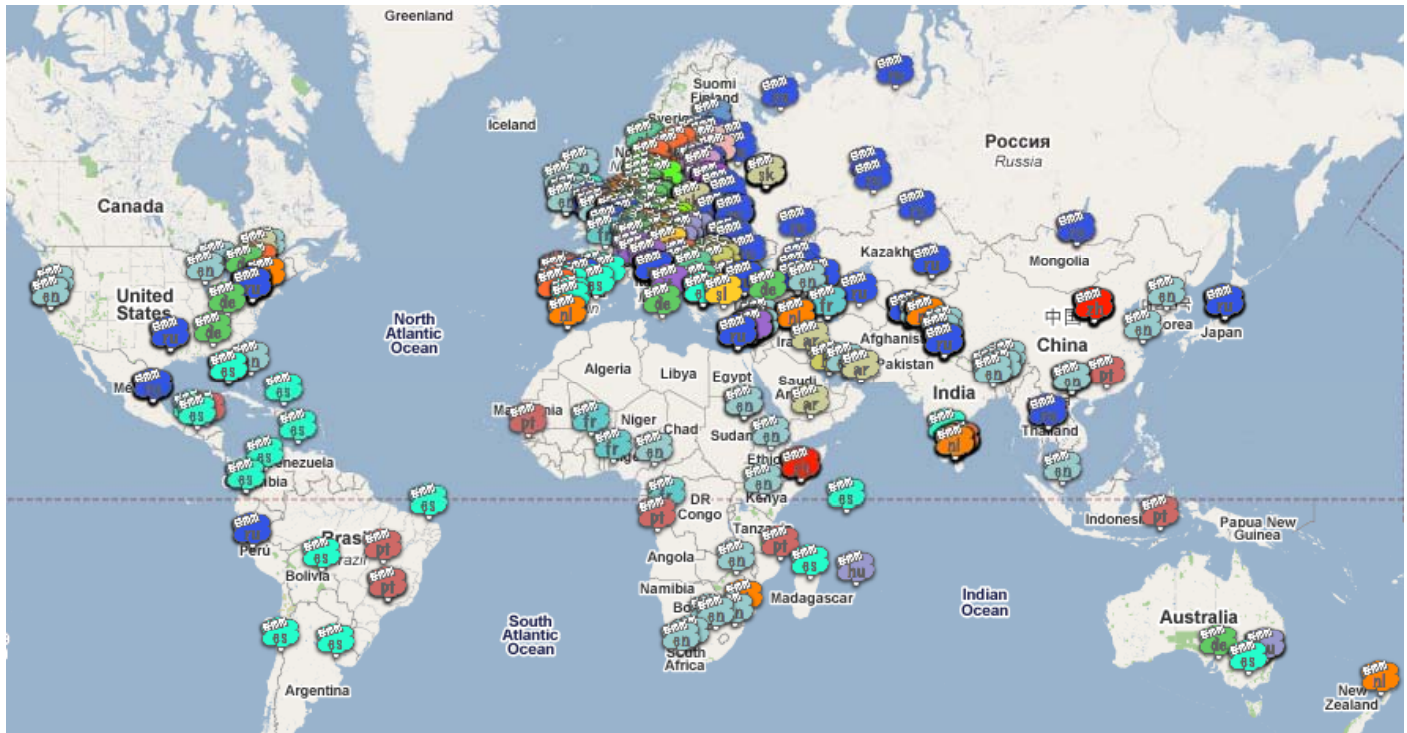


Display of relations between clusters and categories





Display of latest geo-located news clusters



Home Diseases Bioterrorism Nuclear Chemical Other

EMM MedSys
Subscribe to this section RSS RSS feed for this section
Updated every 10 minutes, 24 hours per day.

Search
all - All languages
18 September 2007
20:47:53 o'clock CEST

Latest News - Mustard gas

Mustardgas in the News

Mustardgas Statistics for Mustardgas
15/09/2007 - 18/09/2007

Mustardgas	4
Arsine	1
Chlorine	1
Cyanide	1
Cyanogen chloride	1
Hydrogen	1
Potassium cyanide	1
Sodium	1

Number of articles for this alert | Articles per 10.000 received by the system

Page: 1 2 3 Next

All (61) Medical (0) Newspapers (59) TV/Radio (2) Wires (0)

Page: 1 2 3 Next

[18/09/2007 15:24] (Ξένη δημοσίευση) - ΑΝΑΚΟΙΝΩΣΗ ΤΥΠΟΥ - HELEXPO AE - Απολογισμός 72ης ΔΕΘ
 18 September 2007 14:43:00 o'clock CEST

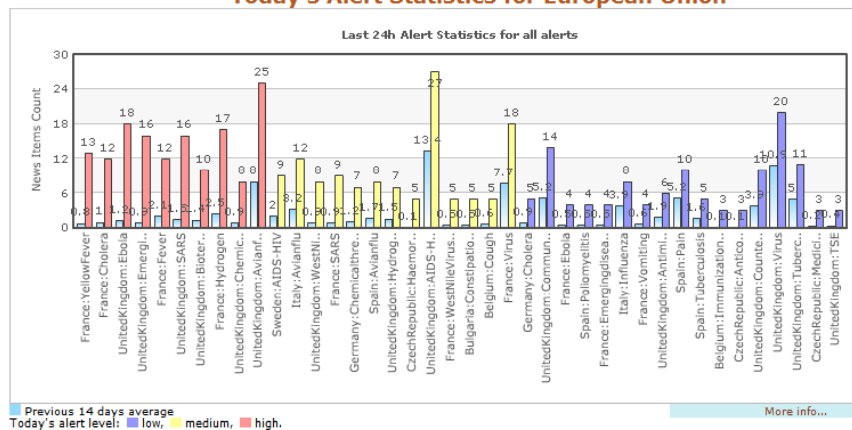
(Ξένη δημοσίευση) - ΑΝΑΚΟΙΝΩΣΗ ΤΥΠΟΥ - HELEXPO AE - Απολογισμός 72ης ΔΕΘ HELEXPO A.E. Γραφείο Τύπου Δελτίο Τύπου (18-9-2007) 72η ΔΕΘ: «ΔΕΘέλουε να τελειώσει»! Αυτή ήταν η εντύπωση - απαίτηση των 250.989 επισκεπτών της 72ης ΔΕΘ που κατέκλυσαν το Διεθνές Εκθεσιακό Κέντρο Θεσσαλονίκης από τις 9

WinGas: Περισσότερο φυσικό αέριο σε Βρετανία και Βέλγιο
 17 September 2007 14:46:00 o'clock CEST

η εταιρεία WinGas, BASF Συγκεκριμένα, η εταιρεία θα αυξήσει φέτος τις προμήθειες στους Βέλγους πελάτες της στα 12,3 δισ. κιολοβάρες αερίου. Για το 2008 έχει ήδη υπογράψει συμφωνίες, οι οποίες

- Documents from all languages get classified according to the same countries and categories.
- An increase of the number of media reports on any country-category combination is detected, independently of the reporting language.
- **Graphs and alerts may show events not yet reported in your own language.**

Today's Alert Statistics for European Union



Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter & Roman Yangarber (2008). **Text Mining from the Web for Medical Intelligence**. In: Fogelman-Soulié Françoise, Domenico Perrotta, Jakub Piskorski & Ralf Steinberger (eds.): Mining Massive Data Sets for Security, pp. 295-310. IOS Press, Amsterdam, The Netherlands

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



EMM **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

EMM-Labs

Освобождены захваченные в турецком банке заложники
news-орел 13.05.2009 17:40:00
Перевести:[ar] [bg] [zh] [hr] [cs] [da] [nl] [en] [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
plan 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....

Google 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

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.



<http://press.jrc.it/NewsExplorer>

NewsExplorer News Analysis

Clustered news for Thursday, September 6, 2007

Luciano Pavarotti 1935-2007 [35]

Italian opera star Luciano Pavarotti dies He was hailed by many as the greatest tenor of his generation....

Analysis over time


Timeline

EMM NewsBrief EMM NewsExplorer

Name Search Text Search

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

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



Countries

- United States (492)
- United Kingdom (196)
- Russian Federation (83)
- Zimbabwe (67)
- France (52)
- China (58)
- Italy (58)
- Afghanistan (56)
- Uganda (54)
- Angola (48)
- South Africa (45)
- Pakistan (42)
- 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)
- Jan 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

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 - Slovensčina
- 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

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 heading 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 waqing aggressive diplomacy against those who

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 president of Cuba and commander-in-chief of Cuba's military 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
شبهه بین المالی فرانسه 24 در برنامه ویژه ای به مناسبت کناره گیری فیدل کاسترو از قدرت در کوبا با تحلیلگری سیاسی خود به گفتگی پرداخت. زمان برنار کادیه تحلیلگری سیاسی این شبکه گفت دوره انتقالی بین از فیدل کاسترو در کوبا از منتهی پیش آغاز شده است. در 31 ژوئیه 2006 وی زمام قدرت را به برادرش راوول 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...
Fidel Castro zrezygnował! de en es fr it nl

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

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

Fidel Castro a anunțat, marti, ca renunța la presedintia Cubei, in editia electronica a cotidianului...
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...
Värmyr 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

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

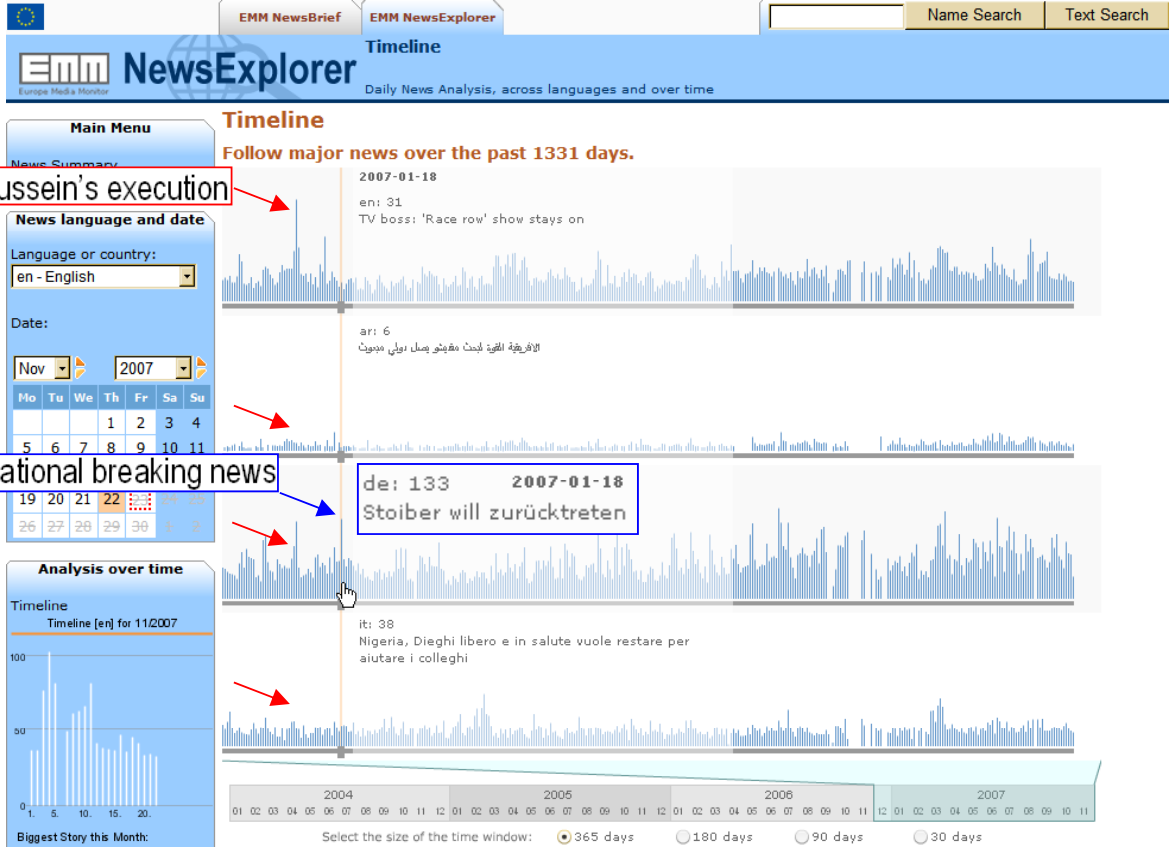
ГАВАНА, 19 февраля. /ИТАР-ТАСС/. Фидель Кастро отказался от поста главы государства и правительства - председателя Государственного совета Кубы. Об этом он сообщил в обращении к...
Bir dönemin sonu de en es fr it nl

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.
hurvetim 10:15:00 CET

Fogh vil ikke savne Castro de en es fr it nl

"Politikberlin" REPLIK: Castro-aja lõpu algus de en es fr it nl

Üks 20. sajandi menukam vabadus-võitleja ja tuntum diktaator Fidel G. kui Nõukogude "geronidid", kelle jaoks tähendas lahkumine võimult ka võim kestab vähemalt esitsotsiaalsed, sest esimeseks asendajaks peeti...
lepl 23:17:00 CET



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

Did Pakistan Face Rule? A Little Test Says No

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

Countries

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

Places

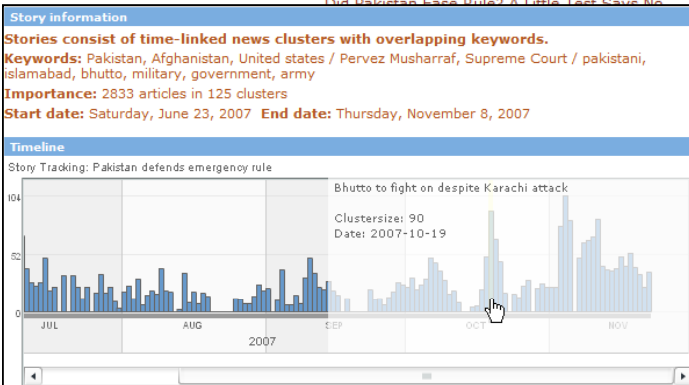
- Islamabad(PK)
- Rawalpindi(PK)
- Karachi(PK)
- Washington(US)
- Ar Riyad(SA)
- Jiddah(SA)

Related Events

- Pervez Musharraf (26)**
- Imran Khan (6)
- George W. Bush (5)
- Gordon Brown (4)
- Don McKinnon (4)
- Nawaz Sharif (4)
- Benazir Bhutto (3)
- Malik Mohammad 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)



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*

NODALIDA-2009, Odense, Denmark, 16 May 2009

Pervez Musharraf

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

Names	Key Titles and Phrases	External resources
Pervez Musharraf (Eu,sv) General Pervez Musharraf (da,sv) Gen Musharraf (en) Pervez Mušaraf (sl) Gen Pervez Musharraf (en) Pervez Musharrafs (da,sv) Pervez Musharraf (da,sv) Первез Мушарраф (ru) Perwez Musharraf (fr,sv) Pervez Moucharraf (fr) برويز مشرف (ar) Perveza Mušarafa (sl) Pervez Muscharraf (de) Pervez Muscharraf (de)	pakistani president (en - 827) president (de,sv - 3230) president gen (en - 506) président pakistanais (fr - 398) pakistaanse president (nl - 235) presidente paquistaniês (pt - 277) pakistani president gen (en - 181) presidente paquistani (es - 147) präsident (de - 617) general (en,sv - 450) præsident (da - 210) presidente (es,pt - 589) president, gen (en - 62)	

Latest Clusters - English

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

Pakistan court dismisses final Musharraf challenge *Struggle to help cyclone survivors in Bangladesh*
euronews 22-NOV-07 *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.
bday 22-NOV-07

[-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]

[said]: The m to improve the instead,

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 Ch 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

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)
- Ha **Other Names**
- Co Al Qaeda (855)
- Wa Supreme Court (505)
- Aft White House (312)
- (2(NATO (207)
- Ayr GEO Television (197)
- Im Lal Masjid (187)
- Abi Tautas Partija (176)
- nt Daily Times (153)

Associated People (51)

- Salman Bashir (1.9)
- Джон Нерпопonte (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

NewsExplorer – Relation exploration

NODALIDA-2009, Odense, Denmark, 16 May 2009

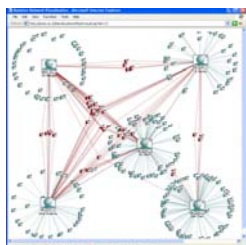
Associated People

- Salman Bashir (1.9)
- Джон Нерпопonte (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)

Example: Pervez Musharraf & Iftikhar Chaudhry



live

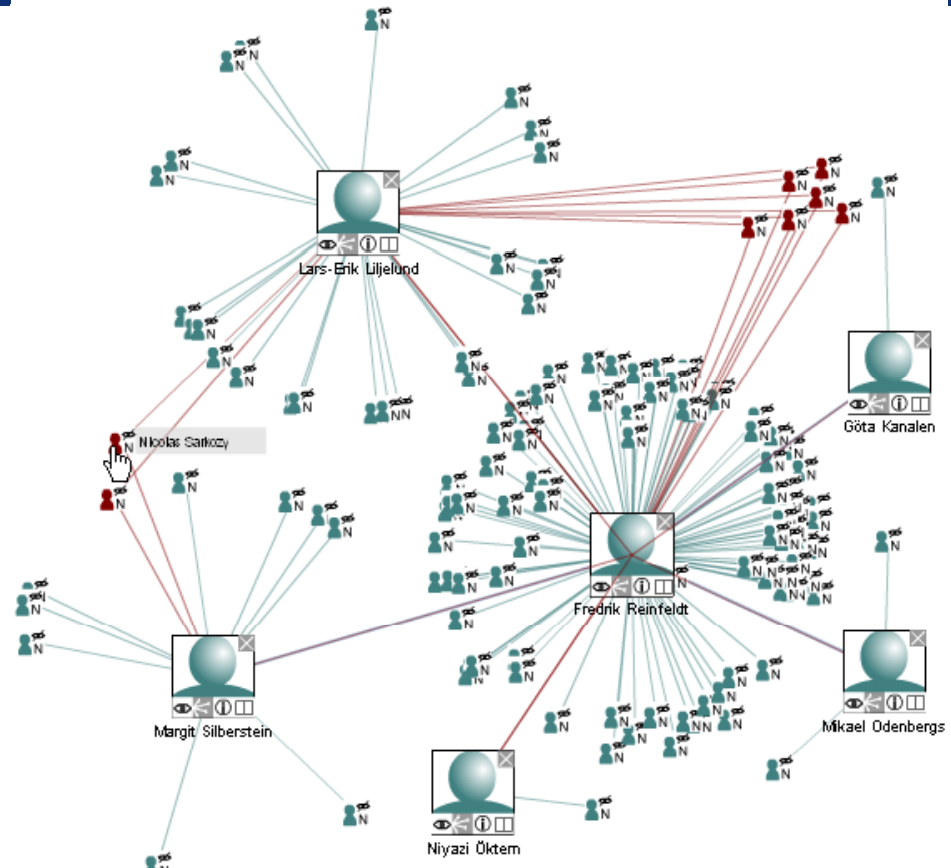


Example:

- [Fredrik Reinfeldt](#)
- Lars-Erik Liljelund
- Margit Silberstein
- ...

Associated People

- Niyazi Öktem (1.2)
- Nicola Clase (1.1)
- Mikael Odenbergs (1.1)
- Margit Silberstein (1.0)
- Göta Kanalen (1.0)
- Gunnar Wetterberg (0.9)
- Bruno Reinfeldt (0.8)
- Mari Ternbo (0.8)
- Lars-Erik Liljelund (0.8)
- Lars-Olov Eriksson (0.8)
- Per-Albin Hanssons (0.8)
- Jan-Åke (0.8)
- Edvard Unsgaard (0.8)
- Anita Kratz (0.7)
- Roberta Alenius (0.7)
- Ulrika Schenström (0.7)
- Hans Heander (0.6)
- Monica Molin (0.6)
- Stig-Björn Ljunggren (0.6)
- Perom Westerbergom (0.6)
- Lennart von Quanten (0.6)
- Martina Krüger (0.6)
- Bjarne Löwdin (0.6)
- Arne Modig (0.6)
- Jan Ertsborn (0.6)
- Anna Hedenmo (0.6)



- Document **clustering**
- • **Named entity recognition** (persons, organisations)
- • **Name variant matching**, including across scripts (transliteration, string similarity calculation)
- **Geo-tagging** (recognition and disambiguation of locations)
- Multi-label document **categorisation**
- **Quotation** recognition (and reference resolution for name parts)
- **Social network generation** (based on co-occurrence, quotations)
- **Topic detection and tracking** (TDT) (Multi-monolingual cluster similarity calculation)
- • **Cross-lingual cluster similarity** calculation (for cross-lingual TDT).

Multilinguality Motivation



Multilinguality: coverage of medical news in various languages

Locations mentioned in MedISys medical articles across languages – **complementary coverage**



Italian - German



English - French



Spanish - Portuguese





Co-occurrence relation between people produced on the basis of many languages is **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)
- Iftekhar 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)

Alexander Litvinenko

Information about this person was last updated on Dienstag, 20. März 2007.

Names	Key Titles and Phrases	External resources
Alexander Litvinenko (Eu,nl)	russo (it,pt - 349)	 <p>Image obtained automatically from Wikipedia</p> <p>Read Wikipedia entry</p>
Alexander Litwinenko (de)	agent russe (fr - 134)	
Alexandre Litvinenko (fr)	ruso (es - 208)	
Aleksandr Litvinenko (fi,no)	agenten (de,sv - 134)	
Aleksander Litvinenko (nl,sv)	kritikers (de - 79)	
Александра Литвиненко (ru)	agent (en,sv - 130)	
Александр Литвиненко (ru)	rusa (it,pt - 76)	
Alexander Livtinenko (it)	agent secret russe (fr - 39)	
Alexander V. Litvinenko (en)	russe (de,fr - 73)	
Alexandr Litvinenko (it)	former russian agent (en - 20)	
Alexander Litvineko (es)	morte di (it - 45)	
Alexandre Livinenko (fr)	ryske agenten (sv - 13)	
Alexander Litvinenko (en)	kritiker (de - 19)	
亞歷山大·利特維年科 (zh)	43 ans (fr - 17)	
Oleksandr Lytvynenko (en)	rusi (it - 14)	
Olexandre Litvinenko (fr)	russian (en - 15)	
Aleksandar Litvinjenko (hr)	omicidio di (it - 11)	
Alexander Litvinenk (it)	officer (en - 13)	
Александр·Литвиненко (ja)	former (en - 16)	
Alexander Walterowitsch Litwinenko (de)		

Steinberger Ralf & Bruno Pouliquen (2007). **Cross-lingual Named Entity Recognition**. In: Satoshi Sekine & Elisabete Ranchhod (eds.), Journal *Linguisticae Investigationes*, Special Issue on Named Entity Recognition and Categorisation, LI 30:1, pp. 135-162. John Benjamins Publishing Company. ISSN 0378-4169.

Multilingual named entity recognition and variant mapping



Names	Key Titles and Phrases
Juan Carlos (da,tr)	rey (es - 686)
Don Juan Carlos (de,pt)	don (es - 616)
rey Juan Carlos (en,es)	könig (de - 327)
Juan Carlos I (de,sv)	rei (pt - 198)
roi Juan Carlos (fr)	rey don (es - 128)
koning Juan Carlos (nl)	king (en - 202)
re Juan Carlos (it)	colombiano (es,pt - 94)
Хуан Карлос (bg,ru)	spanish king (en - 21)
Juan Carlos of Spain (en)	roi (fr - 42)
Juan Carlos da Espanha (Eu,pt)	traficante colombiano (pt - 21)
Juan Carlos Vera (Eu,es)	español (es - 61)
Хуан Карлос I (bg,sr)	re (it - 27)
Juan Carlos Ier (fr)	ingeniero (es - 37)
Juan Carlos I de España (es)	король испании (ru - 14)
Juan Carlos I. (es,sk)	spraanse (nl - 24)
Xoán Carlos I de España (gl)	mari (fr - 23)
Johano Karlo la 1-a (Hispanio) (eo)	obispo (es - 19)
Juan Carlos I na Spáinne (ga)	abogado (es - 36)
könig Juan Carlos (de)	крал (bg - 11)
king Juan Carlos (en)	rey, don (es - 8)
Juan Carlos Ier d'España (ca)	presidente (es,pt - 53)
	ambasciatore (it - 34)



Multilingual name recognition and variant mapping

- en death of former Prime Minister Rafik Hariri, blamed by many opposition
- es asesinato del exprimer ministro Rafic al-Hariri, que la oposición atribuyó
- fr l'assassinat de l'ex-dirigeant Rafic Hariri et le départ du chef de la diplom
- nl na de moord op oud-premier Rafiq al-Hariri gingen gisteren bijna een
- de libanesischen Regierungschef Rafik Hariri vor einem Monat wichtige B
- sl danjega libanonskega premiera Rafika Hariri. Libanonska opozicija si
- et möödumisele ekspeaminister Rafik al-Hariri surma põhjustanud pommipl
- ar اغتيال رئيس الوزراء السابق رفيق الحريري بأيد يهودية وما حدث سابقا
- ru Бывший премьер-министр Ливана Рафик Харири, который

- Lookup of known names from database
 - Currently about 870,000 names
 - + 275.000 variants
 - Only ~167.000 have been found in five different clusters or more
- Pre-generate morphological variants (Slovene example):

Tony(a|o|u|om|em|m|ju|jem|ja)?\s+Blair(a|o|u|om|em|m|ju|jem|ja)

Tony Blair / Tonyju Blairu / Tony Blairt / Tonijs Blērs / Тони Блэра / توني بلير / Tony Blairi / طوني بلير

[Live name variants](#)

Guessing names using empirically-derived *lexical patterns*

- Identification of a current average of 608 unknown names per day
- 83 of those are automatically merged with known names
- **Trigger word(s) + Uppercase Words**
(+ name particles: von, van, de la, abu, bin, ...)
 - President, Minister, Head of State, Sir, American
 - "death of", "[0-9]+-year-old", ...
 - Combinations: "56-year-old former prime minister Kurmanbek Bakiyev"
- Known first names (John, Jean, Hans, Giovanni, Johan, ...)
- Use bootstrapping to produce a trigger word list for a new language
 - Start with small initial trigger word list or lists of known names
 - Produce frequency list of contexts of known names
 - Manual selection

Language Docs/day	Title 1 (pers / occ)	Title 2 (pers / occ)	Title 3 (pers / occ)	Title 4 (pers / occ)	Title 5 (pers / occ)
English 5800	Spokesman (6324/24539)	Dr (5501/11636)	Director (4474/10393)	Mr (4262/11113)	President (4158/63480)
German 2400	Chef (2894/30792)	Deutsche (2148/7577)	Sprecher (1832/6788)	Präsident (1407/23190)	Berliner (1034/3108)
French 1120	président (2688/23629)	M (1172/7658)	Directeur (1129/2754)	Chef (1102/3700)	Français (837/4014)
Spanish 2450	Presidente (2245/21429)	Director (1104/2166)	General (987/3800)	Portavoz (893/2401)	Estadounidense (834/3863)
Portuguese 1450	Presidente (1978/21628)	Segundo (1488/2379)	Diz (955/1252)	Deputado (861/6810)	General (676/3181)
Italian 1080	Presidente (1439/8925)	Avvocato (624/2572)	Ministro (577/4157)	Generale (536/1954)	p.m. (517/2797)
Dutch 1450	Voorzitter (1279/5011)	Directeur (1202/2817)	Woordvoerder (1098/2389)	Minister (726/9954)	President (582/7476)
Swedish 930	VD (CEO) (603/2146)	Talesman (304/753)	Chef (294/603)	Advokat (225/1045)	President (217/3375)
Slovene 400	Minister (438/4624)	Predsednik (417/2652)	Director (260/586)	Predsednika (246/1456)	Dr (245/393)
Estonian 130	Director (169/281)	President (165/1084)	Peaminister (89/676)	Presidendi (77/231)	Direktori (65/90)

(number of persons / number of occurrences)

- Inflection of trigger words for person names, using regular expressions (Slovene example):

- **kandidat**(a|u|om)?
- **legend**(a|e|i|o)
- **milijarder**(ja|ju|jem)?
- **predsednik**(a|u|om|em)?
- **predsednic**(a|e|i|o)
- **ministric**(a|e|i|o)
- **sekretar**(ja|ju|jom|jem)?
- **diktator**(ja|ju|jem)?
- **playboy**(a|u|om|em)?

+ uppercase words

... verskega **voditelja** **Moktade al Sadra** je z notranjim ...
= **Muqtada al-Sadr** (ID=[236](#))



Language	# Rules	# Texts	# Names	Average Precision	Average Recall	Average F-measure
English	1100	100	405	92	84	88
French	1050	103	329	96	95	95
German	2400	100	327	90	96	93
Spanish	580	94	274	85	84	84
Italian	440	100	298	92	90	91
Russian	447	61	157	81	69	74

Steinberger Ralf & Bruno Pouliquen (2007). **Cross-lingual Named Entity Recognition**. In: Satoshi Sekine & Elisabete Ranchhod (eds.), Journal *Linguisticae Investigationes*, Special Issue on Named Entity Recognition and Categorisation, LI 30:1, pp. 135-162. John Benjamins Publishing Company. ISSN 0378-4169.

- Adding names (and images) from Wikipedia
- Merging NewsExplorer name variants
 - Transliteration
 - Normalisation
 - Similarity measure

http://en.wikipedia.org/wiki/Hamid_Karzai

The screenshot shows the Wikipedia page for Hamid Karzai. On the left, there is a list of 'in other languages' with checkboxes for Afrikaans, العربية, Български, Dansk, Deutsch, Eesti, Ελληνικά, Español, Esperanto, فارسی, Français, Gaeilge, Galego, 한국어, हिन्दी, Bahasa Indonesia, Иронay, Italiano, עברית, ქართული, and كوردی / كوردي. The main content area shows the name 'Hamid Karzai' and a brief biography. On the right, there is a photo of Hamid Karzai with the caption 'Hamid Karzai' and 'حامد کرزي'. Below the photo, a list of name variants is shown: Хамид Карзай, Hamid Karzai, Hamid Karzaï, Hamid Karsai, حامد کرزاي, هَامِيدِ كَرزَی, and 哈米德·卡尔扎伊. Arrows point from the language list to the corresponding variant in the list.

- Currently, EMM NewsExplorer transliterates from Arabic, Farsi, Greek, Russian and Bulgarian
- Transliterate each character, or sequence of characters, by a Latin correspondent
 - $\psi \Rightarrow ps$
 - $\lambda \Rightarrow l$
 - $\mu\pi \Rightarrow b$
- Examples of transliterations:
 - Κόφι Ανάν, Greek → Kofi Anan
 - Кофи Аннан, Russian → Kofi Anan
 - Кофи Анан, Bulgarian → Kofi Anan
 - كوفي عنان, Arabic → Kofi Anan
 - कोफी अन्नान, Hindi → Kofi Anan

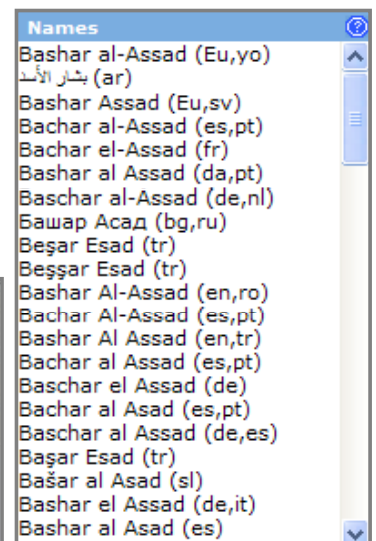
- Transliteration rules depend on the target language, e.g.

Владимир Устинов (Russian)

- **V**ladimir **U**stinov (English)
- **W**ladimir Ustinow (German)
- Vladimir **O**ustinov (French)

- Various ways to represent the same sound: sh, sch, ch, š, e.g.

- Bašar al Assad
- Baschar al Assad
- Bachar al Assad



- Diacritics are often omitted, e.g.
 - Wałesa
 - Walesa
 - Saïd
 - Said
 - Schröder
 - Schroder
 - Skarsgård
 - Skarsgard
 - Jørgen
 - Jorgen

→ Edit distance is large for naturally occurring word variants:

- “Rafik Harriri” vs. “Rafiq Hariri” → 2
- “Rfk Hrr” vs. “Rafiq Hariri” → 6

- Latin normalisation:
 - accented character → non-accented equivalent
 - Malik al-Saïdoullaïev
Malik al-Saidoullaiev
 - double consonant → single consonant
 - Malik al-Saidoullaiev
Malik al-Saidullaiev
 - ou → u
 - Malik al-Saidullaiev
Malik Saidullaiev
 - “al-” →
 - ... mlk sdlv
 - wl (beginning of name) → vl
 - ow (end of name) → ov
 - ck → k
 - ph → f
 - ž → j
 - š → sh
 - x → ks

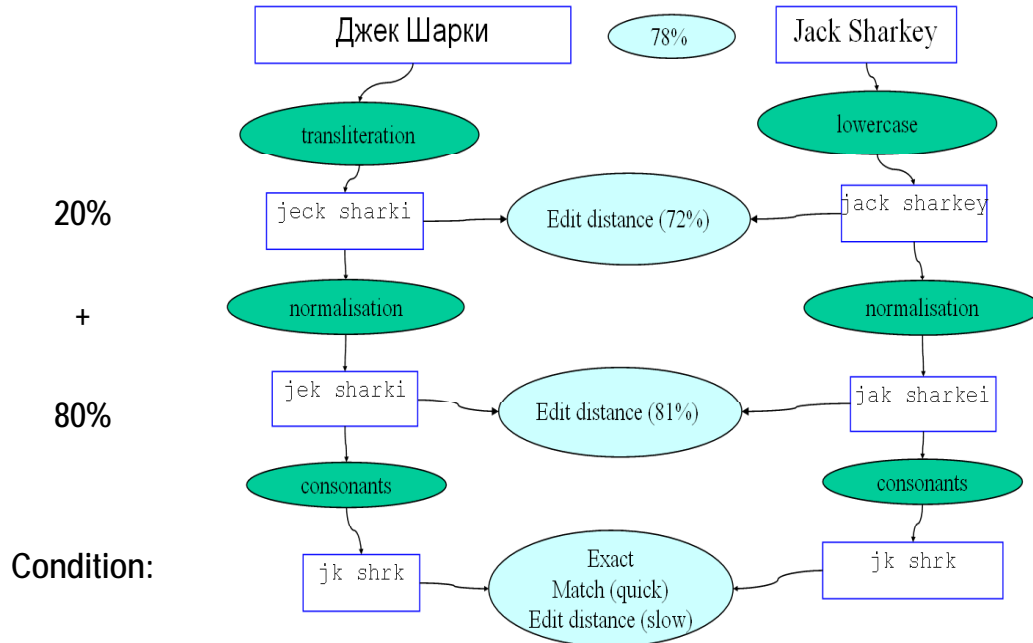
Name	Normalised form
Mohammed Siad Barre, Mohamed Siad Barre, Мохаммед Сиад Барре, محمد سياد بري	mhmd sd br (mohamed siad bare)
Mahmoud Ahmadinejad, Mahmūd Ahmadīnežād	mhmd hmdnjd (mahmud ahmadinejad)
Сергей Куприянов, Sergei Kupriyanov, Sergei Kuprianow, Sergueï Kouprianov	srg kprnv (sergei kuprianov)
Ban Ki-moon, Ban Ki Moon, Пан Ги Мун	bn k mn (ban ki mun)

- Remove vowels

To compare ~600 new names every day with ~1,145,000 known name variants

→ Only if the transliterated, normalised form with vowels removed is identical

- Calculate edit distance variant similarity using two different representations:



- Threshold: 0.94 (100% Precision in test set)
- NewsExplorer: currently, 608 new names every day
 - 83 are automatically merged (14%)
 - Some are additionally saved for expert judgment

Name 1	Name 2	Similarity	Merged?	Same person?
Barzan al-Tikriti	Barzan al Tikriti	0.99	Yes	Yes
Ismail Hanieh	Ismail Hanyieh	0.98	Yes	Yes
Farouq al-Qaddoumi	Farouk al-Kadoumi	0.97	Yes	Yes
Abdullah bin Abdul Aziz	Abdullah bin Abdel Aziz	0.96	Yes	Yes
Barzan al-Tikriti	Barazan al-Takriti	0.94	Yes	Yes
Manfred Wörner	Manfred Werner	0.93	No	No
Michel Ancel	Michael Ancel	0.92	No	Yes
Jorge Costa	Jorge Acosta	0.92	No	No
Falon Gong	Falun Gong	0.90	No	Yes
Roberto Panella	Roberto Pianelli	0.87	No	No
Peter Struck	Peter Starck	0.82	No	No
Jamie Foxx	Jaime Foxx	0.77	No	Yes

Name variants

Ali Larijani (Eu,en)
 Ali Laridschani (de)
 Ali Lariyani (es)
 Ali Laridžani (sl)
 علي لاريجاني (ar)
 Ali Larijani (es)
 Ari Larijani (en)
 Ali Lariyani (es)
 Али Лариджани (ru)
 Ali Larijani (it)
 Ali Laridjani (fr)
 Ali Larjani (sv)
 Ali Lariani (it)
 Ali Laryani (es)
 Ali Laranjani (pt)
 Ali Larigani (en)
 Ali Larinjani (nl)
 Al Larijani (nl)
 Ali Ardashir Larijani (en)
 Ali A. Larijani (en)
 Ali Larejani (it)
 Ali Larichani (es)
 علي اردشير لاريجاني (fa)
 Ali Larijan (en)
 Al Laridschani (de)
 Ali Ardeshir Larijani (en)
 Ali Laridziani (en)



Trigger words

negotiator (en - 824)
 chefunterhändler (de - 470)
 iraniano (it,pt - 311)
 negociador iraniano (pt - 102)
 iranien (fr - 129)
 iranian negotiator (en - 87)
 pogajalec (sl - 90)
 unterhändler (de - 111)
 iraní (es - 121)
 iraanse onderhandelaar (nl -

[live](#)

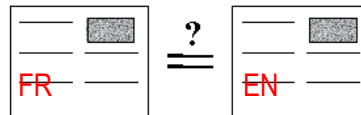


Cross-lingual Cluster Linking

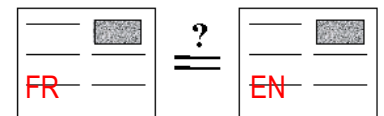
Castro quits as president, state-run
 paper reports [72] de es fr it nl ar bg da
et fa no pl pt ro ru sl sv tr

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





- How to find out whether two texts in different languages are related?
- Most common approach: use MT or bilingual dictionaries to translate into English, then use monolingual methods to calculate similarity.
 - **Using MT** (e.g. Leek et al. 1999 for Chinese-Mandarin to English);
~ 50% performance loss when using MT;
 - **Using bilingual dictionaries** (e.g. Wactlar 1999 for Serbo-Croatian to English; Urizar & Loinaz for Basque, Spanish and English 2007)
 - In TDT 1999, the better results were achieved using MT



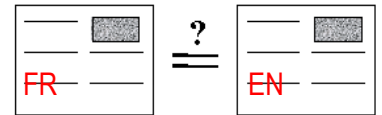
- Automatically produce bilingual lexical space for bilingual document representation and document similarity calculation, e.g.
 - Bilingual *Lexical Semantic Analysis (LSA)* (Landauer & Littman 1991)
 - *Kernel Canonical Correlation Analysis (KCCA)* (Vinokourov et al., 2002)
- + Achieved results are relatively good
- Bilingual approach is restricted to a few languages (OK for English as target lang.):

$$\text{Language pairs} = N * (N-1) / 2$$

(N = number of languages)

- EU: 22 official languages → 231 language pairs (462 language pair directions)!

- Alternative: use entities and thesauri as anchors:
 - Names of persons and organisations
 - Names of locations
 - Dates
 - Terms from multilingual specialist dictionaries (MeSH for medicine, etc.)
 - ...
- Normalise these expressions



→ Use as kind of an interlingua; no language pair-specific resource needed

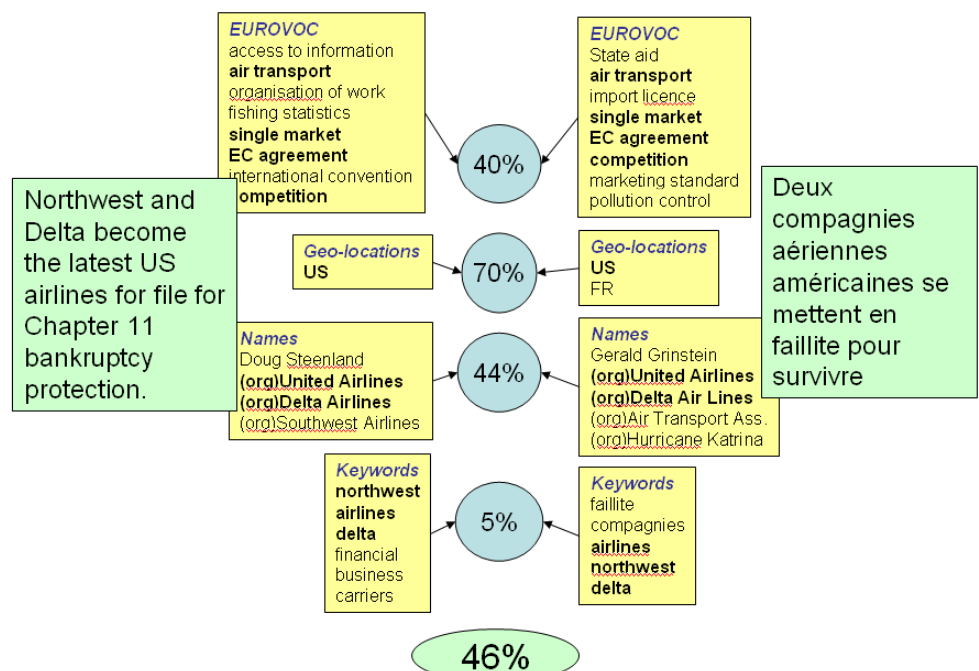
Steinberger Ralf, Bruno Pouliquen & Camelia Ignat (2004). **Providing cross-lingual information access with knowledge-poor methods**. In: Andrej Brodnik, Matjaž Gams & Ian Munro (eds.): *Informatica*. An international Journal of Computing and Informatics. Vol. 28-4, pp. 415-423. Special Issue 'Information Society in 2004'. ISSN: 0350-5596. The Slovene Society Informatika, Ljubljana, Slovenia.

Language-independent features for multilingual document representation

No MT or bilingual dictionaries

19 languages

$$CLDS = \alpha \cdot S1 + \beta \cdot S2 + \gamma \cdot S3 + \delta \cdot S4$$





- **Aim:** recognise geographical references → put a dot on a map
- Lookup of place names from a gazetteer
- **Challenges:**
 - Multilingual gazetteer:
 - combination of multiple sources
 - Inflection → similar to lookup of known persons
 - Homography
 - places-places
 - places-persons
 - places-words
- Usage of various heuristics for disambiguation

English	
Place name	Country
And	Iran
To	Ghana
Be	India
By	Sweden
Are	Nigeria
This	France
But	Afghanistan
Had	Oman
She	India
We	Zaire

Place name	Nb. of cities with this name
Aleksandrovka	244
San Antonio	205
Santa Rosa	199
...	
San Francisco	102
Buenos Aires	88
Washington	32
London	18
Berlin	15
Paris	15
Rome	15
Moscow	12

Name	City: Country
Tony Blair	<i>Tony</i> : USA
	<i>Blair</i> : Malawi
Kofi Annan	<i>Kofi</i> : Mali
	<i>Annan</i> : Scotland
Javier Solana	<i>Javier</i> : Spain
	<i>Solana</i> : Philippines

Pouliquen Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuat, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best (2006). **Geocoding multilingual texts: Recognition, Disambiguation and Visualisation**. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 53-58. Genoa, Italy, 24-26 May 2006. (PDF)



- **eurovoc** THESAURUS
 - ~ 6,000 subject domains
 - Exists in one-to-one translations in all official EU languages, and more
 - Used by many European parliaments for *manual* classification → use for training classifiers
- **Challenges:**
 - Concepts rather than words, e.g. PROTECTION OF MINORITIES, CONSTRUCTION AND TOWN PLANNING
 - Large number of classes (~ 6000)
 - Very unevenly distributed
 - Various text types (heterogeneous training set)
 - Multi-label categorisation (both for training and assignment)
- → **Profile-based category ranking** task (Supervised Learning)
 - Training: Identification of most significant words for each class
 - Assignment: combination of measures to calculate similarity between profiles and new document
- **Result:** Long, weighted list of (numerical) subject domain identifiers → language-independent

Bruno Pouliquen, Steinberger Ralf, Camelia Ignat (2003). **Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus**. In: Proceedings of the Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology - Its Potential and Practicalities (EUROLAN'2003).

Title: Legislative **resolution** embodying Parliament's opinion on the proposal for a Council Regulation amending Regulation No 2847/93 **establishing a control system applicable to the common fisheries policy** (COM(95)0256 - C4-0272/95 - 95/ 0146(CNS)) (Consultation procedure)

Descriptor ID	Descriptor text	Cosine
f 5641040706000000	FISHING CONTROLS [g]	0.360
f 5641020000000000	FISHING GROUNDS [nt]	0.308
f 5641040200000000	COMMON FISHERIES POLICY [g]	0.280
f 5641040100000000	FISHERY MANAGEMENT [nt]	0.279
f 5641040700000000	FISHING REGULATIONS [g]	0.270
f 5641040704000000	FISHING PERMIT [g]	0.261
f 5641040101000000	CONSERVATION OF FISH STOCKS [s]	0.253
f 5641040600000000	FISHING AREA [g]	0.252
f 5206040100000000	CONSERVATION OF RESOURCES [s]	0.251
f 5641050000000000	FISHERY RESOURCES	0.232
f 5641040800000000	CATCH OF FISH	0.213
f 5641040000000000	FISHERIES POLICY	0.203
f 5641040705000000	FISHING LICENCE	0.181
f 5641060100000000	FISHING FLEET	0.179
f 5641010000000000	FISHING INDUSTRY	0.176
f 5641040201000000	EUROPECHE	0.176

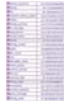
- Freely available for research purposes on our web site:
- <http://langtech.jrc.it/JRC-Acquis.html>
- Total of over one Billion words
- Pair-wise alignment for all 231 language pairs!
- Most documents have been Eurovoc-classified manually

Language-independent features for multilingual document representation

No MT or bilingual dictionaries

19 languages

$$CLDS = \alpha \cdot S1 + \beta \cdot S2 + \gamma \cdot S3 + \delta \cdot S4$$



Sim1 (40%):
Multilingual Eurovoc
subject domains

10.4184	"us"
1.5610	"gb"
1.5610	"fr"
1.5610	"br"

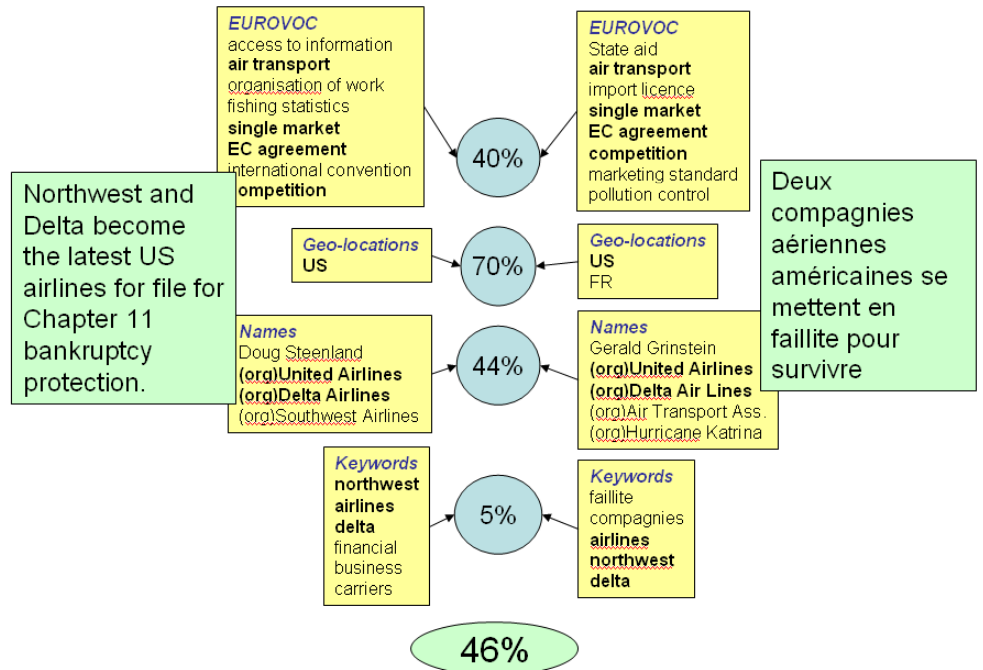
Sim2 (30%):
Geo-locations



Sim3 (20%):
Names + variants

Keynes	Keynes
100.2470	jackson
41.5400	marshall
27.2947	smith
32.6100	monetarism
24.2100	my
24.4211	gdp
20.8824	documentary
18.7973	aircover
13.2940	courthouse
11.1244	jay
10.0838	rich
8.8001	multinational

Sim4 (10%):
Cognates and numbers
(without country score)



Summary and ongoing work

- EMM family of applications
 - Functionality of NewsExplorer
- Technical details on selected NewsExplorer components:
 - Person and organisation name recognition (NER)
 - Name variant mapping
 - Cross-lingual cluster linking
- Highly multilingual applications → simplistic methods
- Each and every component of NewsExplorer could and should be improved!

Ralf Steinberger, Pouliquen Bruno & 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, Amsterdam, The Netherlands.

- Ongoing activities:
 - Re-implementing various tools in Java
 - Transfer more NewsExplorer functionality to the other EMM applications
 - Blog monitoring
- Current research:
 - Learning rules of machine transliteration of names
 - Sentiment analysis
 - Multilingual multi-document summarisation