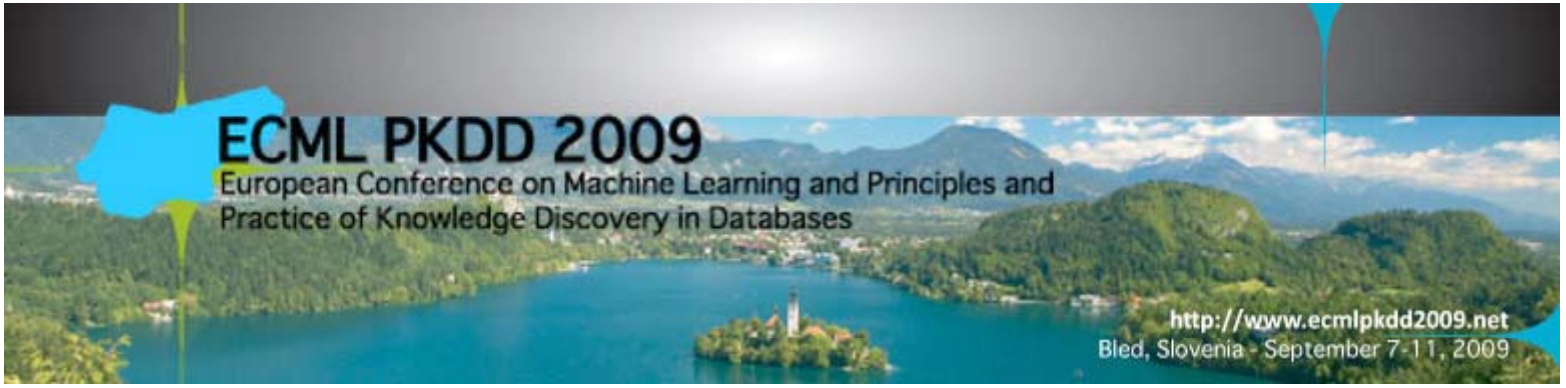
EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Language Technology

Highly multilingual news monitoring applications

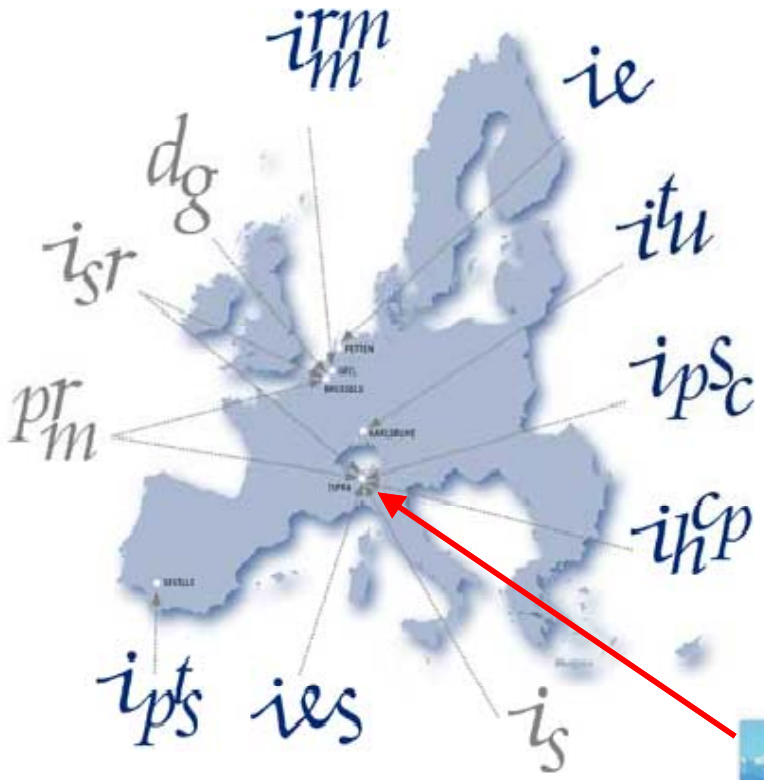


Ralf Steinberger

& 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**)

Agenda

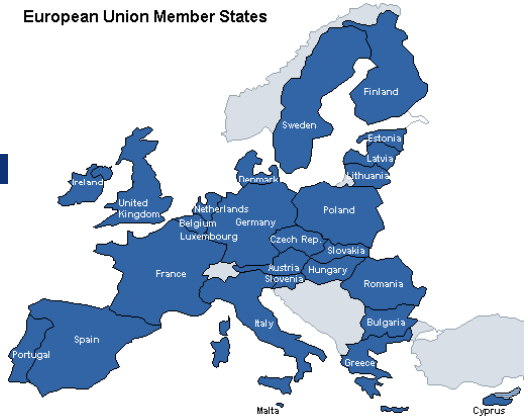
- Media Monitoring and multilinguality
- Europe Media Monitor (EMM) applications - Functionality
  - Publicly accessible at http://press.jrc.it/overview.html

- Some language technology components in detail
  - Multilingual person name recognition
  - Name variant matching across many languages
  - Social networks based on multilingual information extraction
  - Cross-lingual cluster linking
- Summary and Ongoing work

Media Monitoring – Users / Motivation

- Any large organisation wants to monitor
  - newest developments in their field of interest
  - their public image, people's views on certain issues, etc.
  - …
- Institutions and governments, e.g. EC, UN, National Public Health institutions
  - Security-related events (bombings, hostage takings, piracy, …)
  - Humanitarian events (displaced people, shortage of water/medicines, …)
  - Threats to public health (contagious diseases, nuclear or chemical leaks, …)
  - …
- The public (journalists, specialists, anybody)
  - General information / following the news
  - Meta-news sites are independent of individual (possibly biased) news sources

European context – Motivation for multilinguality

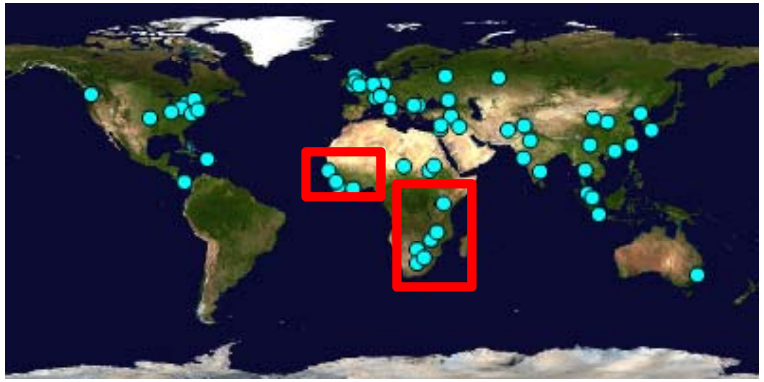


- The European Union has 23 official languages,
  - Plus other, non-official ones
  - Plus an interest in non-EU languages (Turkish, Croatian, Arabic, Russian, Farsi, Chinese, …)
- Political and practical need to cover many languages
- Ideally no pivot language, but direct information access between the language pairs
- The EC's Translation Service DGT is one of the largest in the world (2350 staff plus freelancers, see http://ec.europa.eu/dgs/translation/index_en.htm)
- Complementary coverage of the news in different languages

Multilinguality: complementary news coverage in various languages

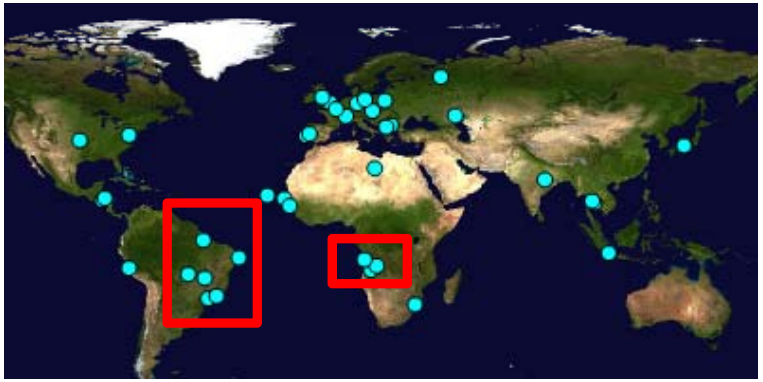
Locations mentioned in MedISys medical articles across languages – complementary coverage

Italian - German



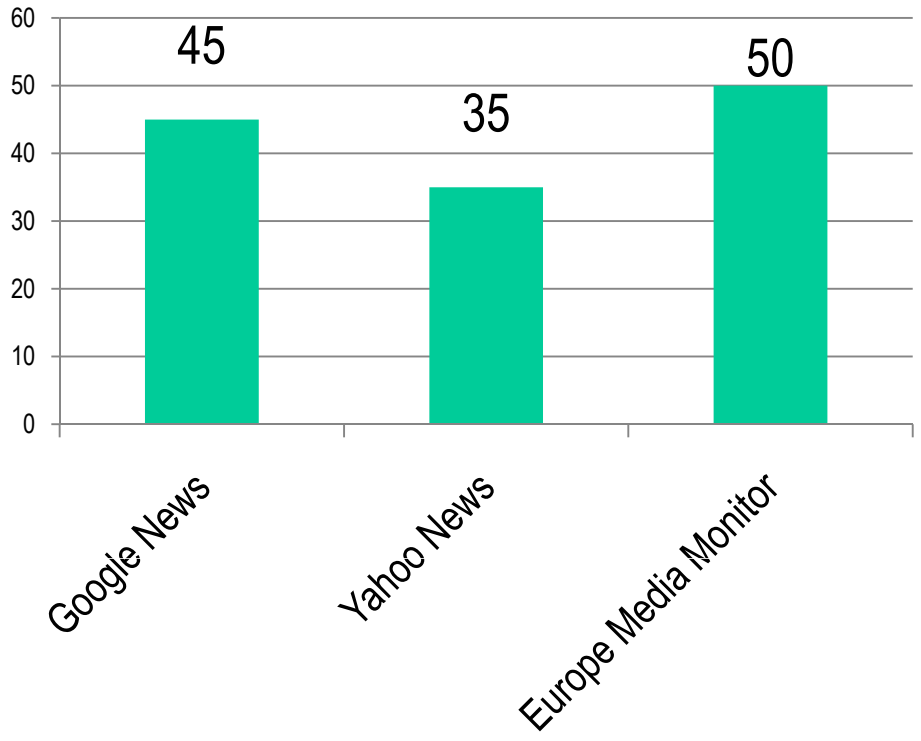
English - French

Spanish - Portuguese



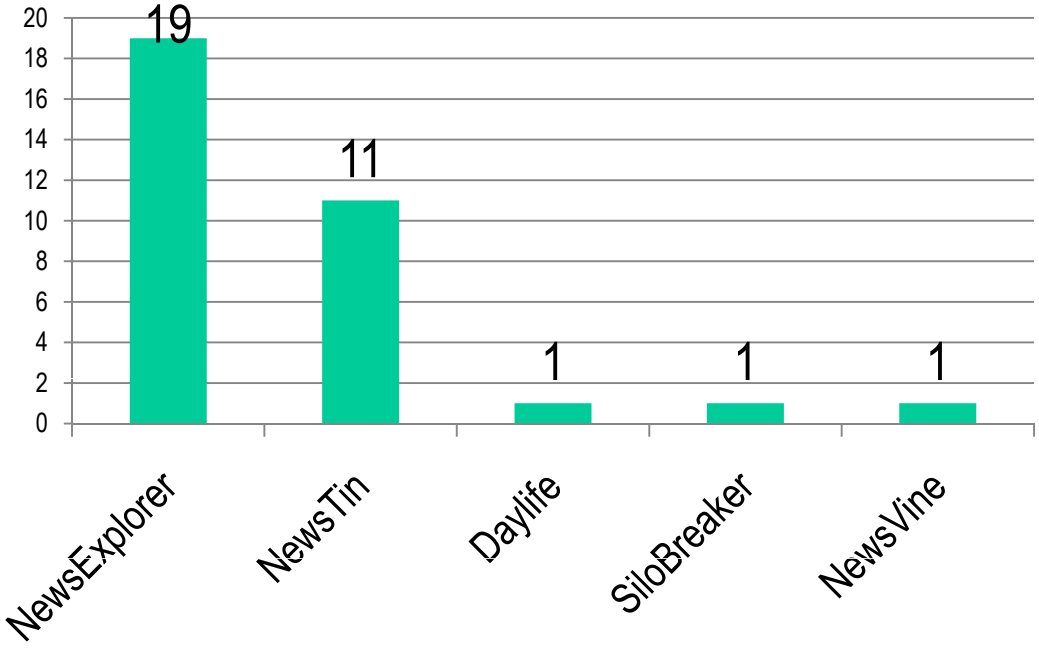
Language coverage of various media analysis tools (August 2009)

News aggregators



in early 2008: 34, 17, 43 languages

News analysis systems



in early 2008: the same

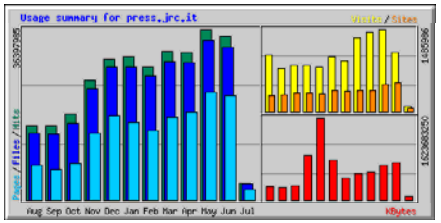
Agenda

- Media Monitoring and multilinguality
- Europe Media Monitor (EMM) applications - Functionality
  - Publicly accessible at http://press.jrc.it/overview.html

- Some language technology components in detail
  - Multilingual person name recognition
  - Name variant matching across many languages
  - Social networks based on multilingual information extraction
  - Cross-lingual cluster linking, incl. multi-label categorisation using Eurovoc
- Summary and Ongoing work

Europe Media Monitor (EMM) - News data

- EMM news gathering engine
  - Monitors ~ 2,200 news sources
  - Gathers 80,000 – 100,000 news articles per day
  - In about 50 languages
  - Visits some sites every 5 minutes
  - Extracts text from the web page
  - Converts text into Unicode-encoded RSS
  - Feeds the news into the four publicly accessible media monitoring systems

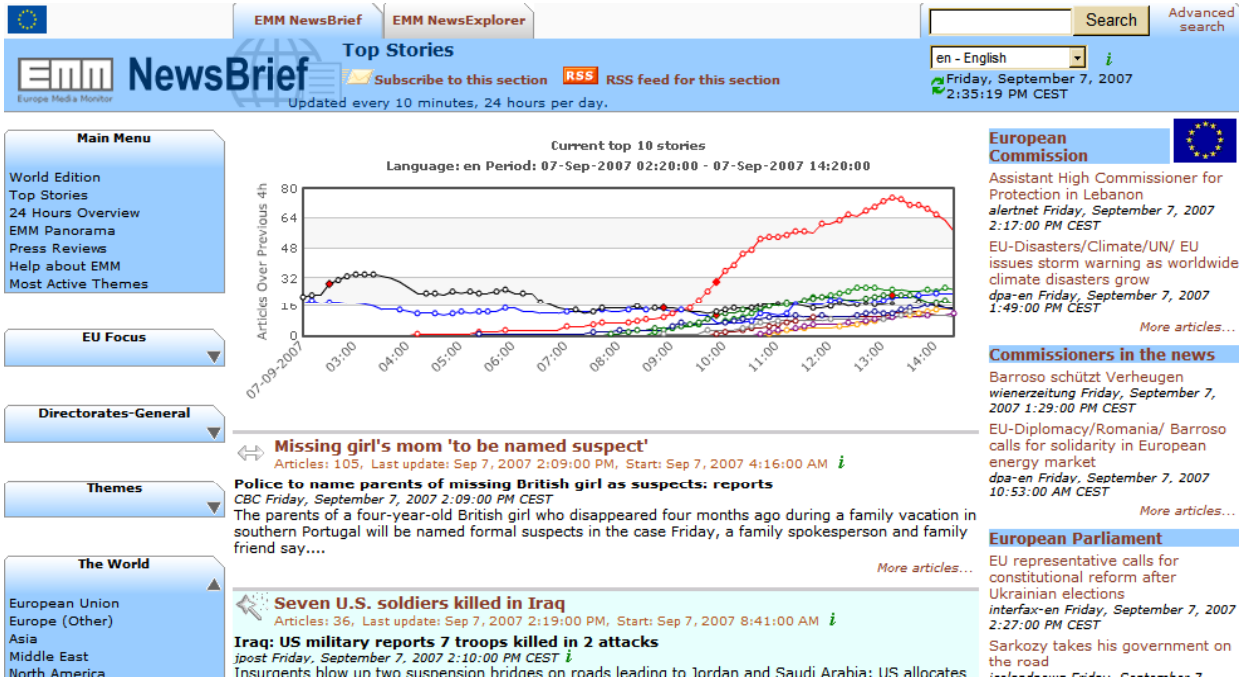


- Combined between 1 and 2 Million hits per day
- 30,000 – 50,000 distinct users per day

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.

EMM – NewsBrief (currently 50 languages)

- Public site: http://press.jrc.it/NewsBrief/ (since 2002)
- Categorises news into ~ 600 categories, using:
  - Boolean search word combinations
  - vicinity operators
  - optional weights
  - regular expressions
- Clusters and tracks news *live* (multi-monolingually)
- Detects breaking news
- Sends out email notifications for each category

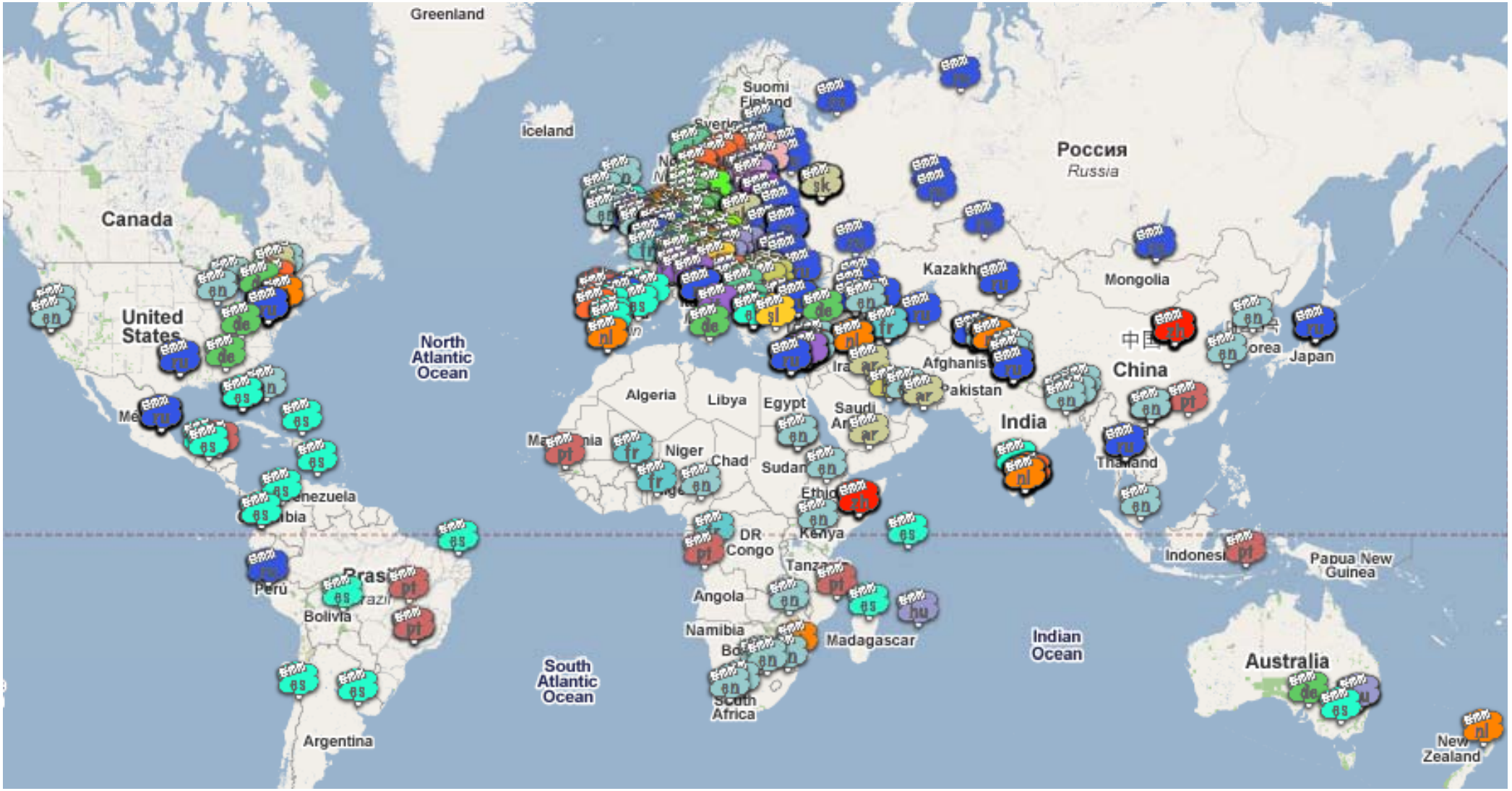


Steinberger Ralf, Bruno Pouliquen & Erik van der Goot (2009). **An Introduction to the Europe Media Monitor Family of Applications**. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8.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

JRC
EUROPEAN COMMISSION

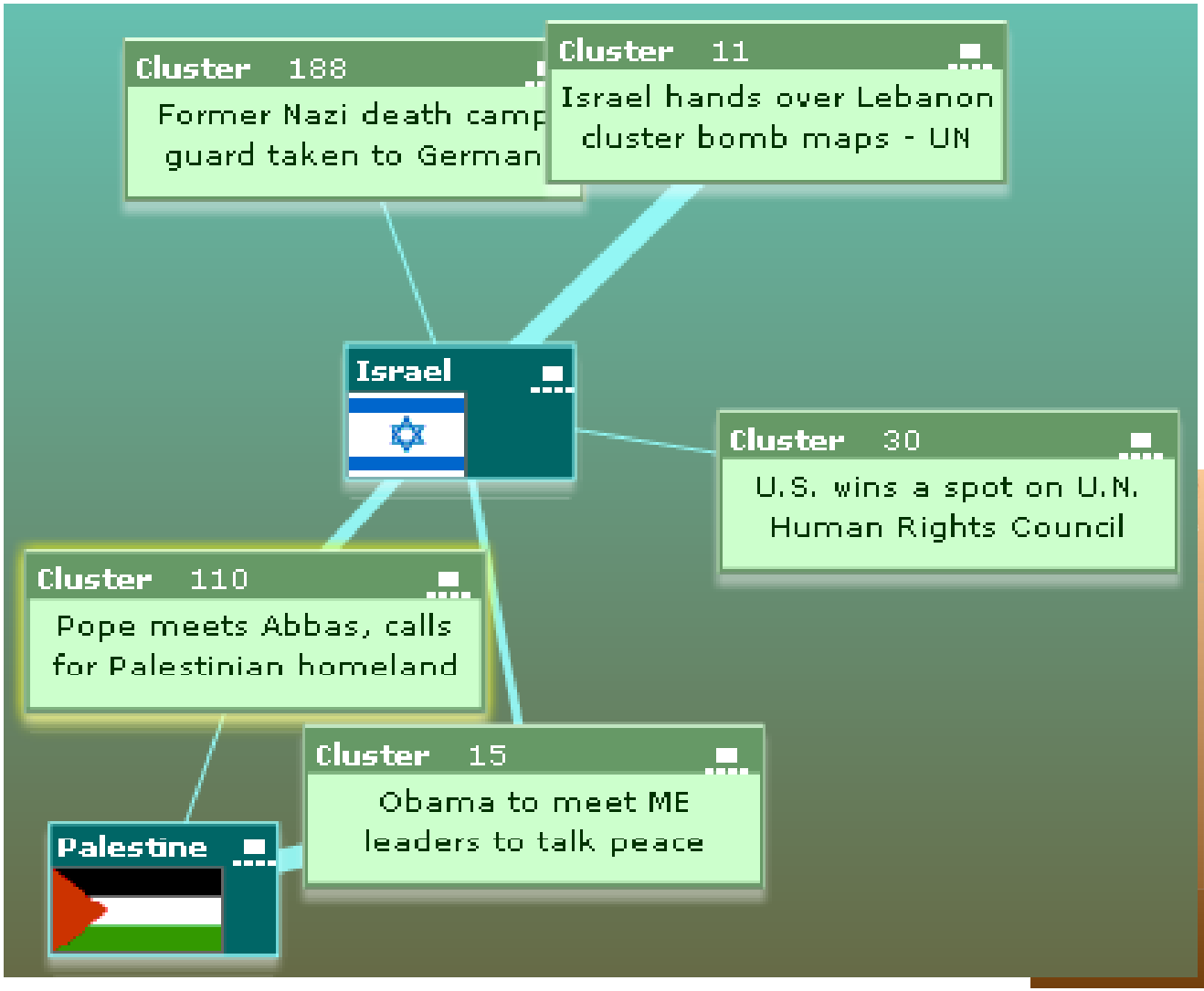
NewsBrief: Live Cluster Map

Live display of latest geo-located news clusters



EMM-Labs: Real-time Cluster Network

Display of relations between clusters and categories



MedISys – Filtering and classification in up to 50 languages

Home | Diseases | Bioterrorism | Nuclear | Chemical | Other

Search | Advanced search

EMM MedISys Medical Information System — Europe Media Monitor

Mustard gas

Subscribe to this section | RSS feed for this section

Updated every 10 minutes, 24 hours per day.

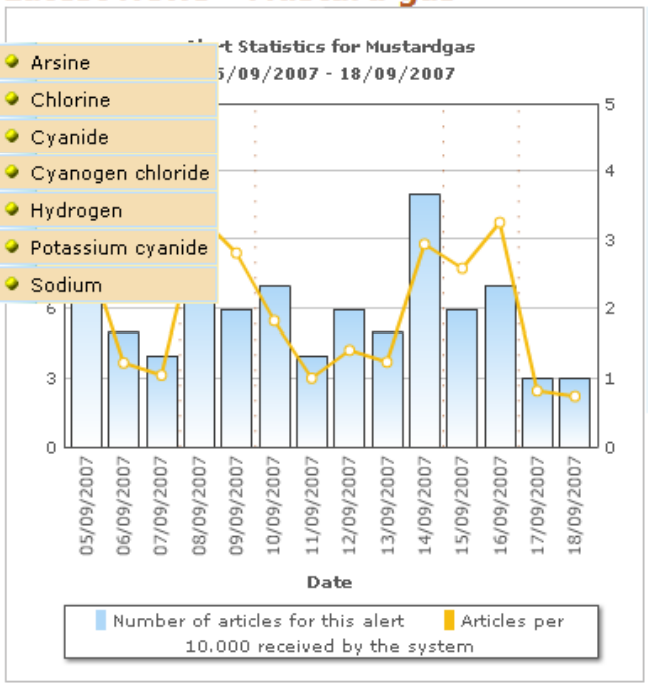
all - All languages

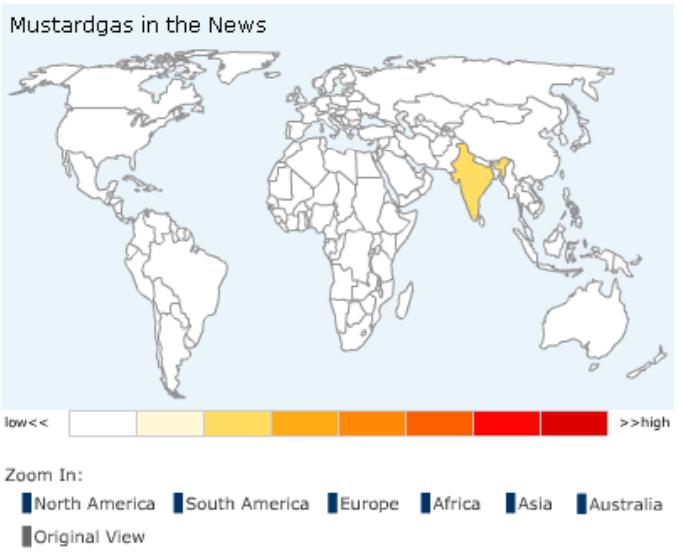
18 September 2007
20:47:53 o'clock CEST

- chemical
- Blood agents
  - Arsine
  - Chlorine
  - Cyanide
  - Cyanogen chloride
  - Hydrogen
  - Potassium cyanide
  - Sodium
- Blister agents
- Nerve agents
- General terms
- Metals
- Vomiting agents
- Riot control agents
- Incapacitating agents
- Choking-Pulmonaryagents
- Anticoagulants
- Other agents

Latest News - Mustard gas

Page: 1 2 3 Next





The country values are calculated by setting the number of articles mentioning a theme AND a country in relation to the number of articles about the theme and the number of articles about the country.

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

ana 18 September 2007 14:43:00 o'clock CEST

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

أسعار "الإيثيلين جلايكول" تواصل الصعود وسط استقرار لفالبية المنتجات وسابك تؤكد التزامها بإمدادات الأوكسجين

alriyadh 18 September 2007 03:51:00 o'clock CEST

استقر مؤشر أرقام البتروكيماويات، والذي يقيس حركة أسعار البتروكيماويات لسلة من المواد التي يتم إنتاجها في الخليج العربي، عند مستوى 252.9نقطة منهيا فترة تسعة أسابيع من الارتفاع المتتالي ، مدفوعا بالمستويات

(إسرائيل) تزعم إحباط هجوم على مستعمرة.. والاعتقالات تطال 13بالضفة

alriyadh 18 September 2007 03:51:00 o'clock CEST

استشهد فتى في السادسة عشرة من العمر خلال توغل لقوات الاحتلال في رام الله صباح أمس، فيما زعمت سلطات الاحتلال احباط عملية اقتحام لمستعمرة "شافي شومرون" شمال غربي نابلس....

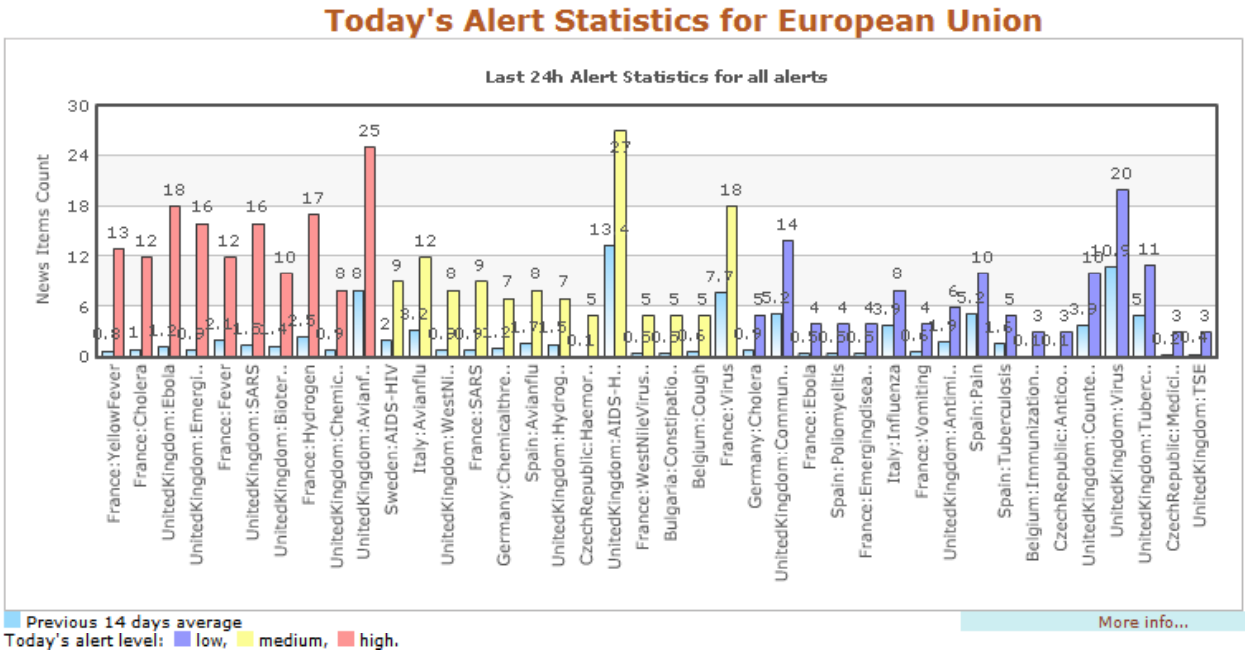
WinGas: Περισσότερο φυσικό αέριο σε Βρετανία και Βέλγιο

naftemporiki 17 September 2007 14:46:00 o'clock CEST

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

Aggregation of multilingual information

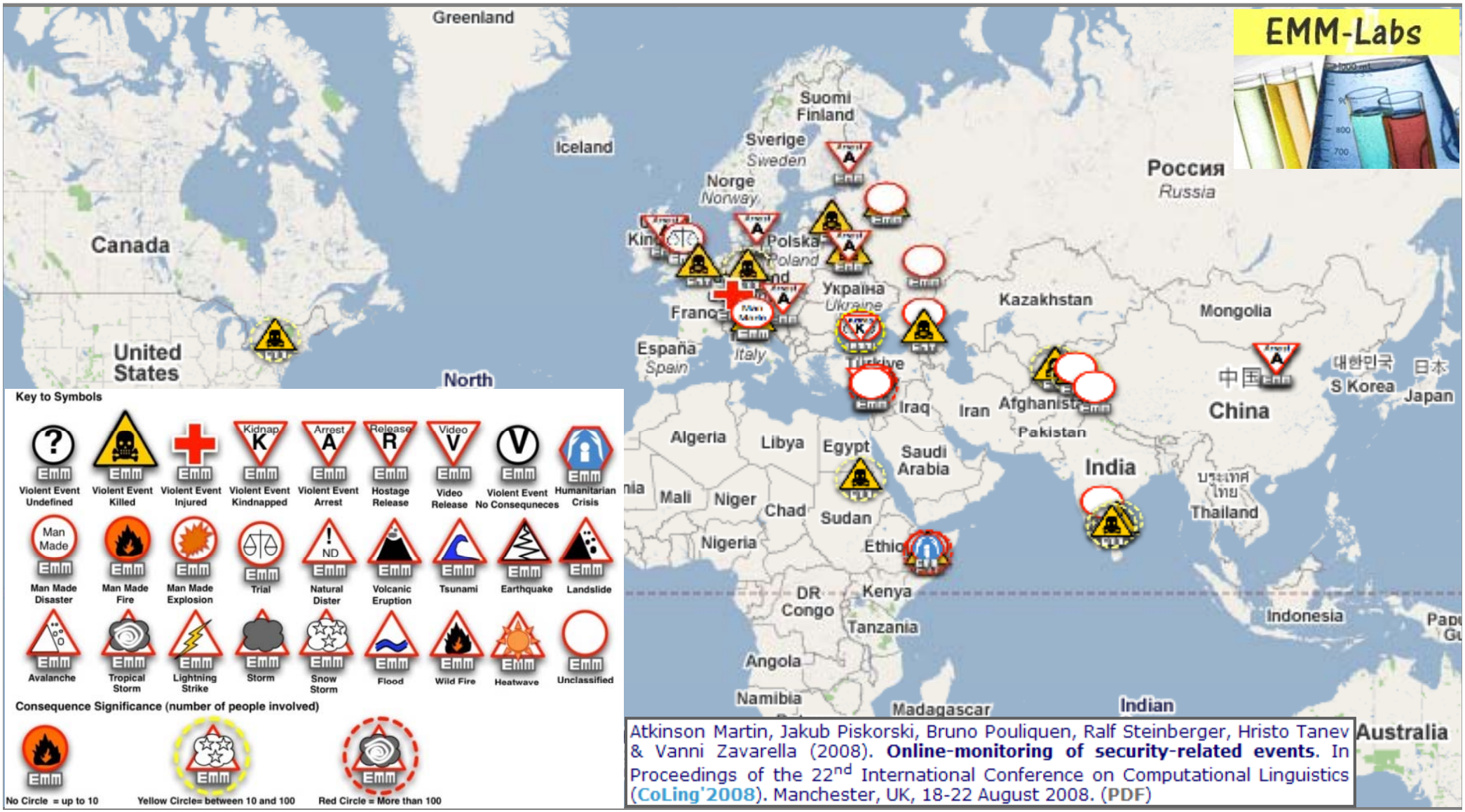
- Documents from all languages get classified according to the same countries and categories.
- An increase of the number of media reports on any country-category combination is detected,
- independently of the reporting language.
- Graphs and alerts may show events not yet reported in your own language.



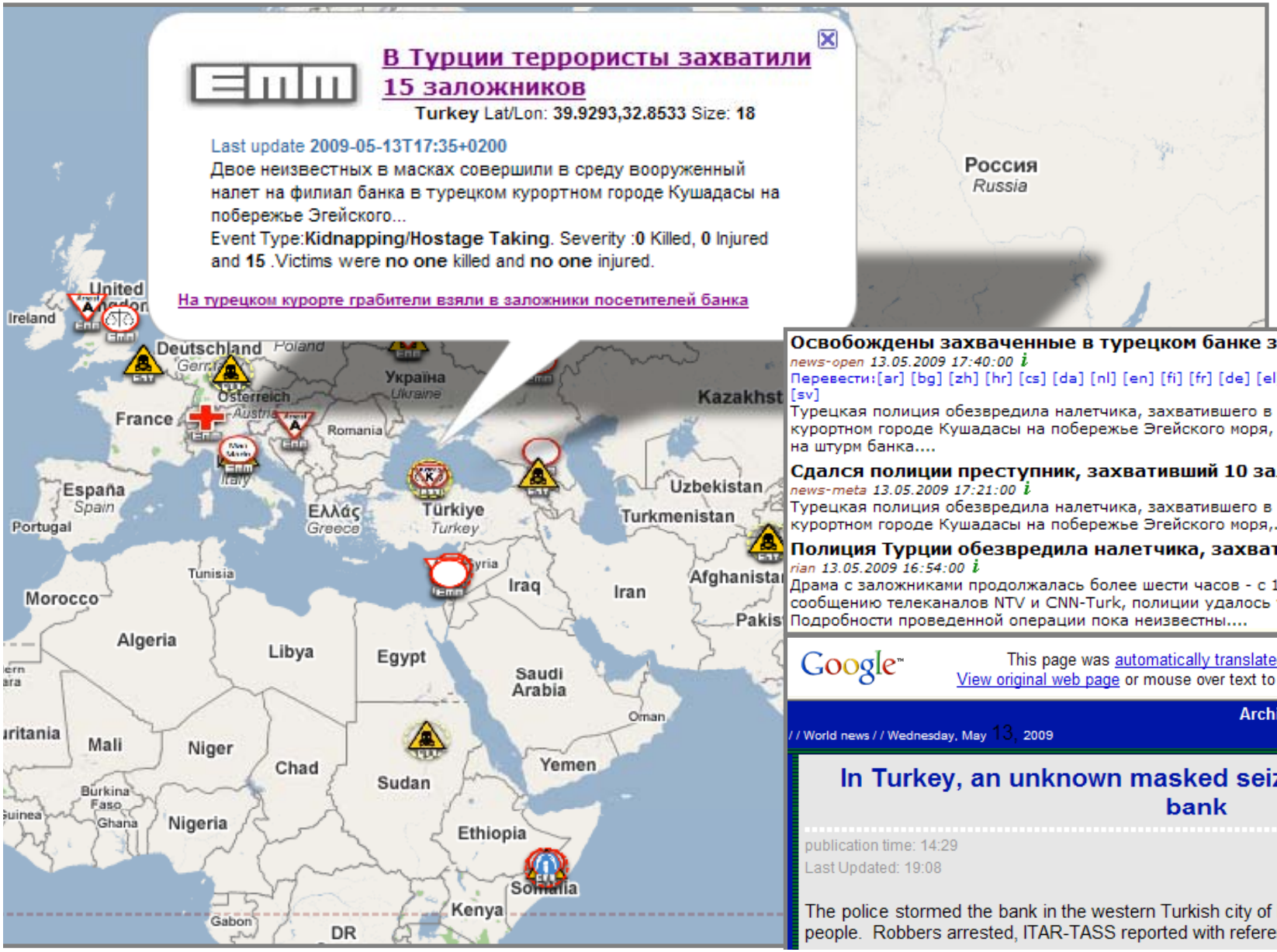
Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter & Roman Yangarber (2008). **Text Mining from the Web for Medical Intelligence**. In: Fogelman-Soulié Françoise, Domenico Perrotta, Jakub Piskorski & Ralf Steinberger (eds.): Mining Massive Data Sets for Security. pp. 295-310.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

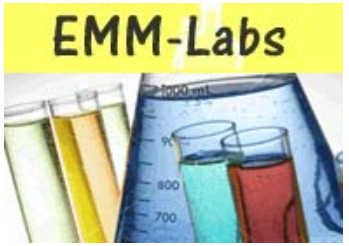
Detection and visualisation of events (violence/disasters/humanitarian/…)

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



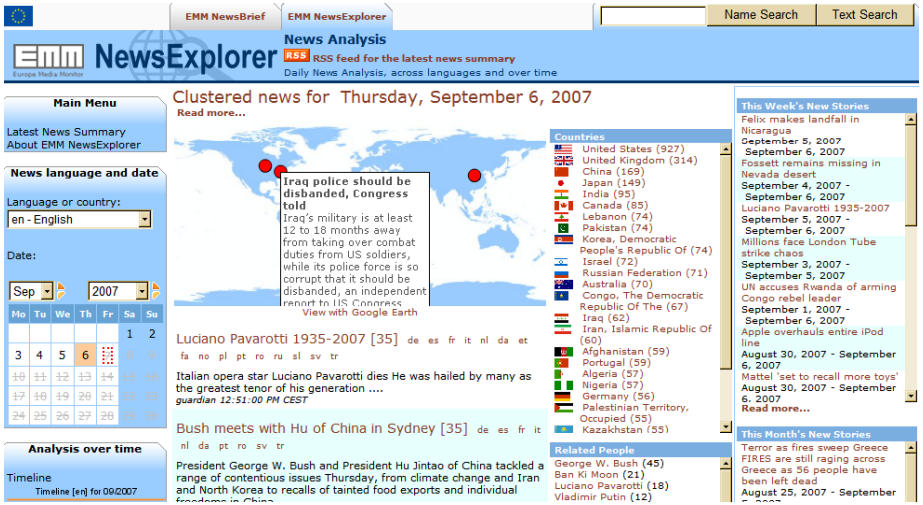
Detection and visualisation of events (violence/disasters/humanitarian/…)





NewsExplorer – Functionality

http://press.jrc.it/NewsExplorer



Since April 2004

NewsExplorer – Multilingual daily news overview

EMM NewsBrief | EMM NewsExplorer | Name Search | Text Search

EMM NewsExplorer — Europe Media Monitor

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

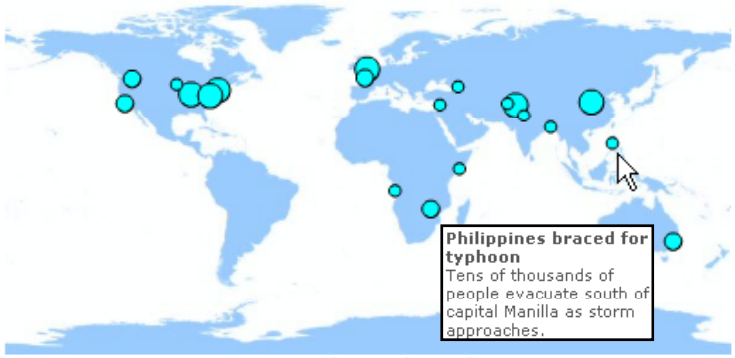
ar - Arabic
bg - български
es - Español
de - Deutsch
da - Dansk
en - English
et - Eesti keel
fa - Farsi
fr - Français
it - Italiano
nl - Nederlands
no - Norsk
pl - Polski
pt - Português
ro - Română
ru - Russian
sl - Slovenščina
sv - Svenska
tr - Türkçe

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...



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

NewsExplorer – Cross-lingual cluster linking

ECML-PKDD, Bled, Slovenia, 9 September 2009

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 رائول 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 går av de en es fr it nl

Partiav
som pr
styrtet
stat på
VG

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 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 ru sl sv tr

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 ويوش يدعو للتحول الديمقراطي في كوبا de en fr it nl

(CNN) -- هافانا، كوبا في مؤتمر صحفي في رواندا، إلى مساعدة كوبا على البدء بعملية "انتقال ديمقراطية"، وذلك إثر قرار الزعيم 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

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

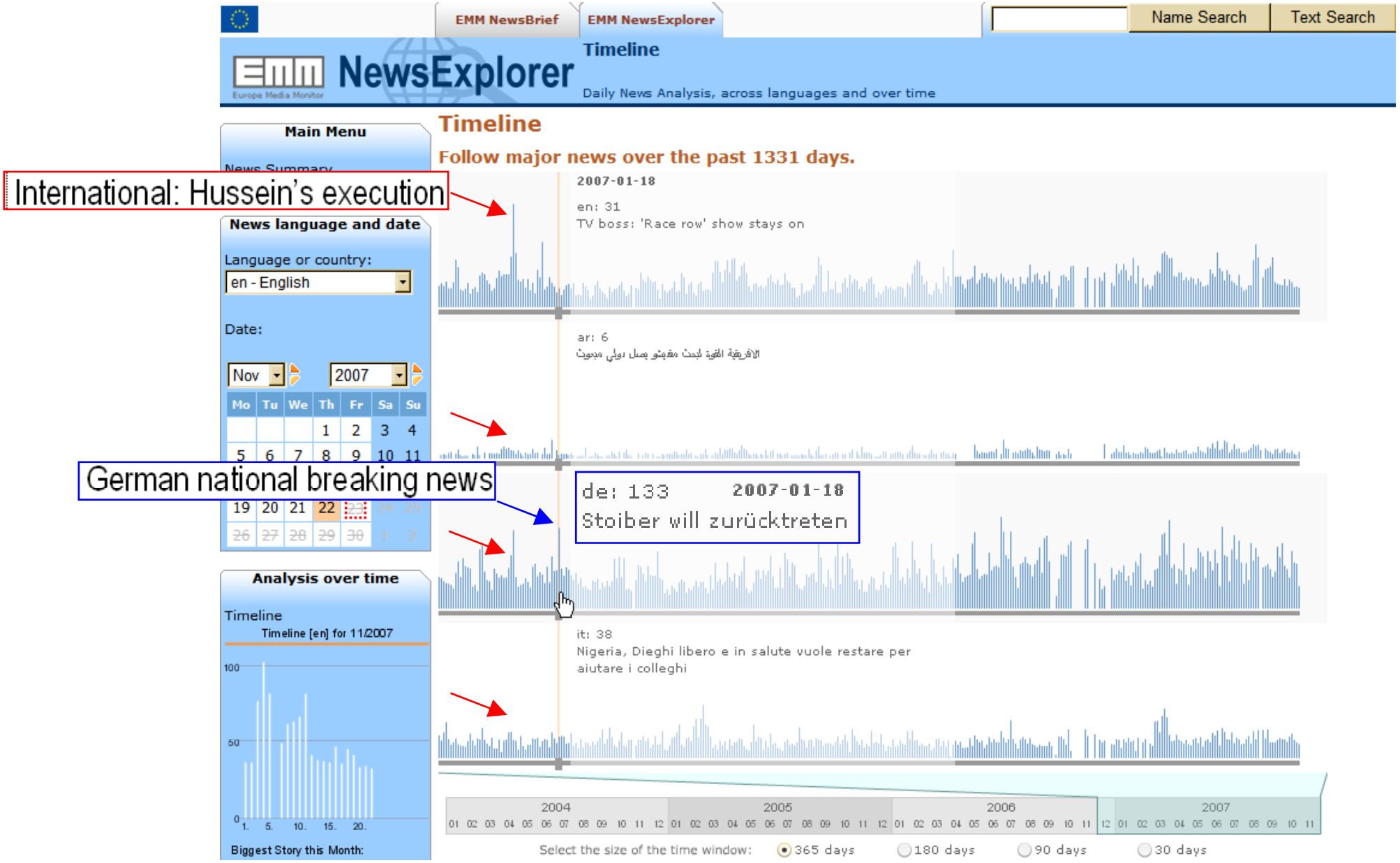
Bir dönemin sonu de en es fr it nl

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

NewsExplorer – Time line: biggest clusters per day



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

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 People
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)

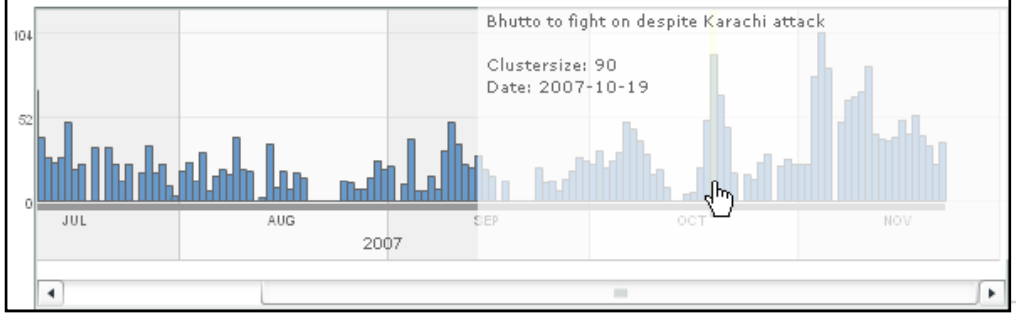
Other Names
Supreme Court (27)
Human Rights Watch (3)
Pakistan Muslim League (2)
High Court (2)
High Commission (2)
GEO Television (2)
Human Rights Commission (1)
Al Qaeda (1)

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



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

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 Мушарраф (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)

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

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

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]

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

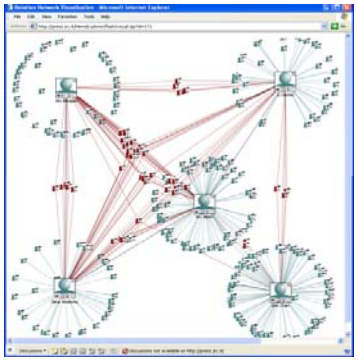
NewsExplorer – Relation exploration

live

Example: Pervez Musharraf & Iftikhar Chaudhry

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)
- Qazi Hussain Ahmed (0.9)



Text analysis tools integrated in NewsExplorer

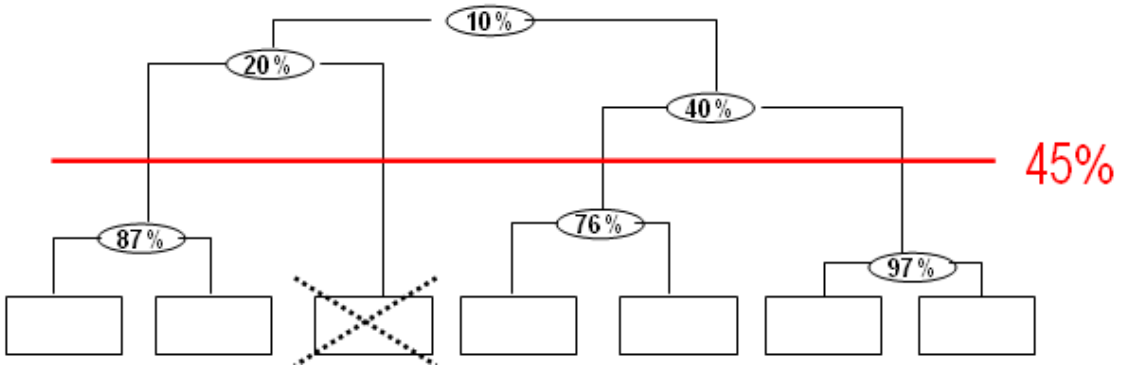
- Document clustering
- Named entity recognition (persons, organisations)
- Name variant matching, including across scripts (transliteration, string similarity calculation)
- Geo-tagging (recognition and disambiguation of locations)
- Multi-label document categorisation
- Quotation recognition (and reference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

Agenda

- Media Monitoring and multilinguality
- Europe Media Monitor (EMM) applications - Functionality
  - Publicly accessible at http://press.jrc.it/overview.html

- Some language technology components in detail
  - Multilingual person name recognition
  - Name variant matching across many languages
  - Social networks based on multilingual information extraction
  - Cross-lingual cluster linking, incl. multi-label categorisation using Eurovoc
- Summary and Ongoing work

News clustering

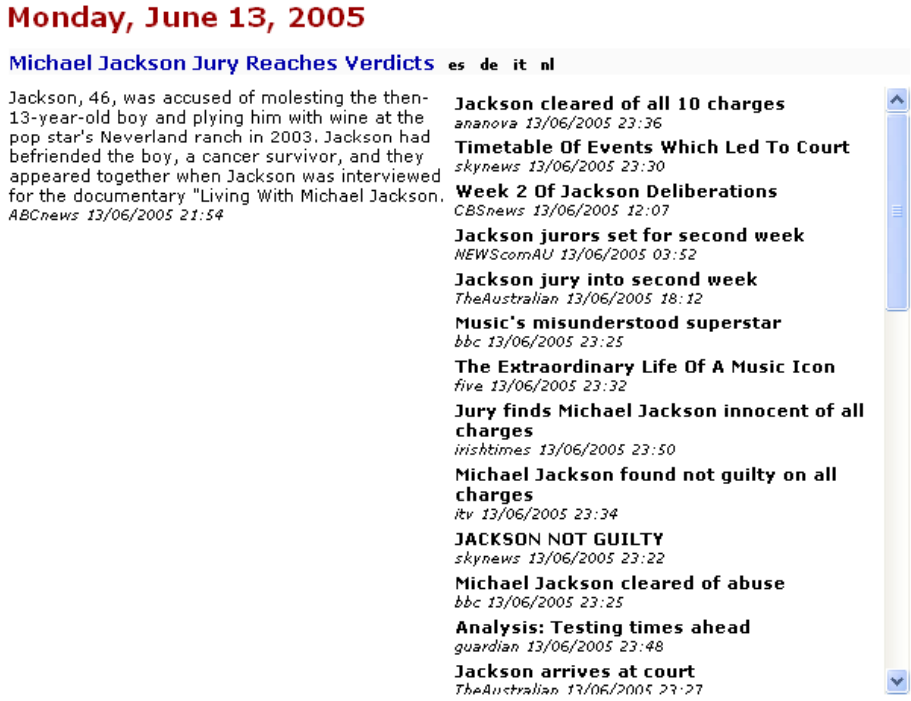


Clustering: Monolingual document representation

- Vector of keywords and their keyness using log-likelihood test (Dunning 1993)
- Enhanced with log-likelihood weighted country score

"Michael Jackson Jury Reaches Verdicts"

Keyness	Keyword	Keyness	Keyword
109.2478	jackson	7.5620	testimony
41.5450	neverland	6.5014	maria
37.9347	santa	4.0957	michael
32.6105	molestation	1.7368	reached
24.5193	boy	1.6857	ap
24.4351	pop	1.5610	*gb*
20.6824	documentary	1.5610	*il*
18.7973	accuser	1.5610	*br*
13.5945	courthouse	1.0520	appeared
11.1224	jury	0.5384	child
10.4184	*us*	0.5045	trial
10.0838	ranch	0.4502	monday
9.6021	california	0.2647	children
9.3905	verdict	0.0946	family

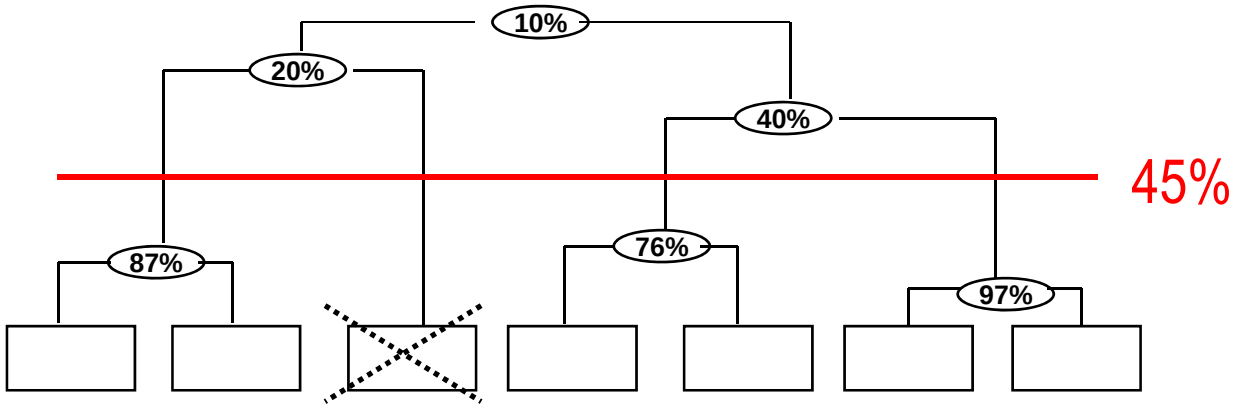


Original cluster

Multi-monolingual news clustering

- Document similarity measure: cosine
- Method: Bottom-up group average unsupervised clustering
- Build the binary hierarchical clustering tree (dendrogram)
  - Retain only "big" nodes in the tree with a high cohesion (empirically refined minimum intra-node similarity: 45%)
- Use the title of the cluster's medoid as the cluster title

Keyness	Keyword
109.2478	jackson
41.5450	neverland
37.9347	santa
32.6105	molestation
24.5193	boy
24.4351	pop
20.6824	documentary
18.7973	accuser
13.5945	courthouse
11.1224	jury
10.4184	*us*
10.0838	ranch
9.6021	california
9.3905	verdict



Pouliquen Bruno, Ralf Steinberger, Camelia Ignat, Emilia Käsper & Irina Temnikova (2004). **Multilingual and cross-lingual news topic tracking**. In: Proceedings of the 20th International Conference on Computational Linguistics (CoLing'2004), Vol. II, pages 959-965. Geneva, Switzerland, 23-27 August 2004. (PDF)

NewsExplorer

Multilingual named entity recognition and variant mapping

Names	Key Titles and Phrases
Barack Obama (Eu,yo)	prezydent usa (pl - 189)
Barak Obama (az,wo)	senator (da,sv - 322)
Барак Обама (ba,uk)	sen (en - 332)
باراك أوباما (ar)	prezydent (pl - 141)
باراك اوباما (ar,fa)	senador (es,pt - 123)
باراک اوباما (fa)	us-präsident (de - 40)
Barack Hussein Obama (da,tr)	abd başkanı (tr - 30)
Barrack Obama (da,tr)	senador democrata (pt - 35)
Obama Barack (da,sv)	prezydenta (pl - 46)
Barack Hussein Obama II (an,tt)	rival (da,fr - 58)
Barack Hüseyin Obama (tr)	candidate (en - 66)
Barac Obama (en,tr)	president (de,sv - 127)
Barack Hussain Obama (en)	presidentskandidaat (nl - 27)
Barrack Hussein Obama (en,fr)	sénateur (fr - 38)
Barrak Obama (en,tr)	amerykański prezydent (pl - 14)
Барак Хуссейн Обама (ru)	ameriškega predsednika (sl - 19)
Baraque Obama (pt)	democrat (en - 35)
バラク・オバマ (ja)	amerikaanse president (nl - 18)
Barack Hussein Obama Jr. (ca)	
באראק אובאמא (yi)	

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 Haririja. Libanonska opozicija si

et möödumisele ekspeaminister Rafik al-Hariri surma põhjustanud pommipl

ar اغتيال رئيس الوزراء السابق رفيق الحريري بأياد يهودية وما حدث سابقا

ru Бывший премьер-министр Ливана Рафик Харири, который

Multilingual named entity recognition (NER) – brief overview

- Lookup of ~1.15 million known names and variants from database
  - Currently about 870,000 names (status June 2009) + 275.000 variants
  - Dealing with morphological variation by pre-generating 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)

- Guessing new names using empirically-derived *lexical patterns*
  - President, Minister, Head of State, Sir, American
  - "death of", "[0-9]+-year-old", …
  - Known first names + uppercase words
  - Identification of a current average of 608 unknown names per day

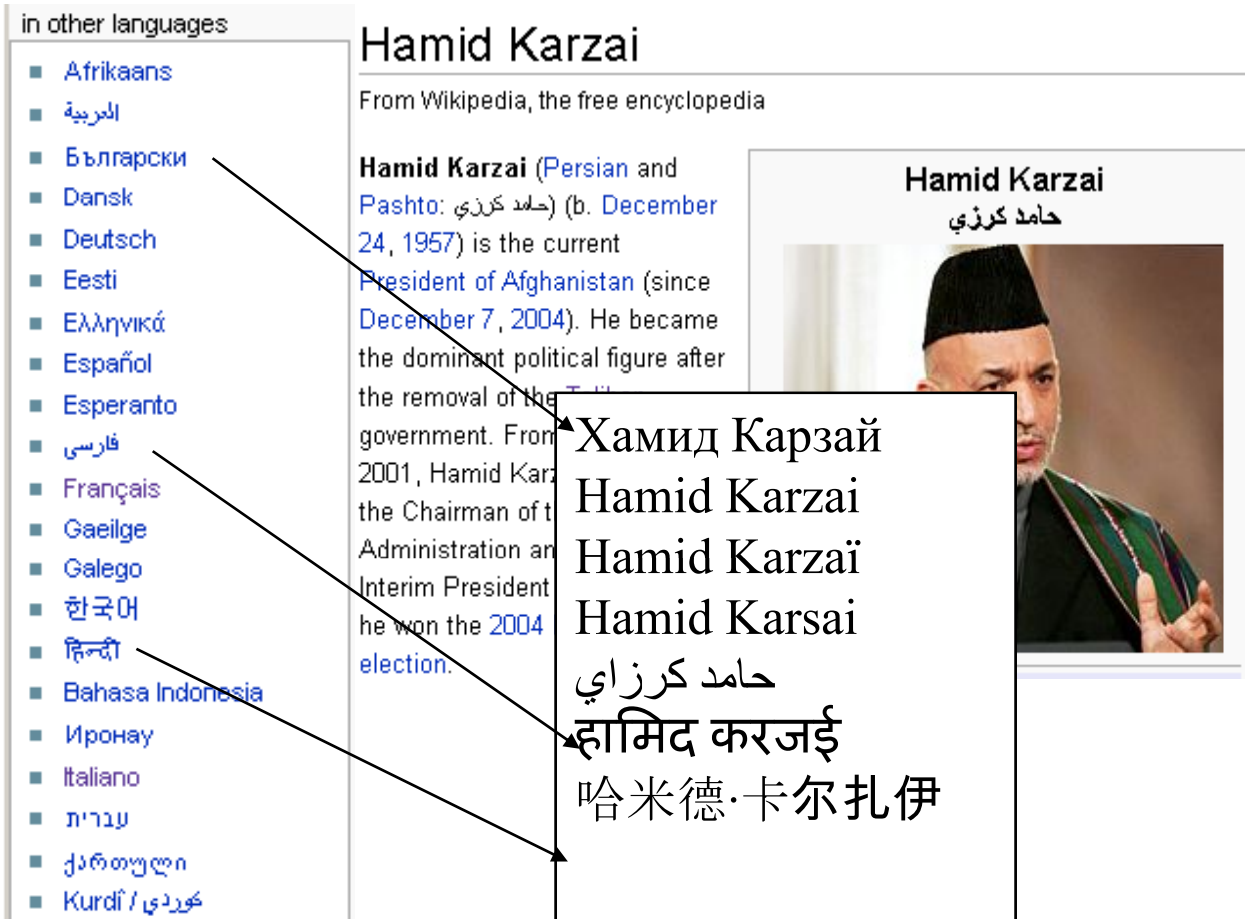
Steinberger Ralf & Bruno Pouliquen (2007). **Cross-lingual Named Entity Recognition.** In: Satoshi Sekine & Elisabete Ranchhod (eds.), Journal **Linguisticae Investigationes**, Special Issue on Named Entity Recognition and Categorisation, LI 30:1, pp. 135-162. John Benjamins Publishing Company. ISSN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

Live name variants

Name Variants

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

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



Name transliteration

- Currently, EMM NewsExplorer transliterates from
  Arabic, Farsi, Greek, Russian and Bulgarian

- Transliterate each character, or sequence of characters, by a Latin correspondent
  - ψ => ps
  - λ => l
  - μπ => b

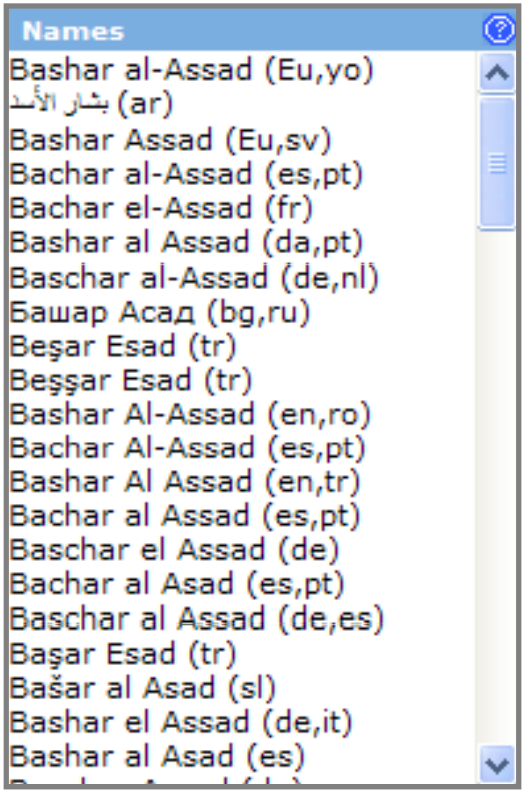
  - Examples of transliterations:
    - Κόφι Ανάν, Greek → Kofi Anan
    - Кофи Аннан, Russian → Kofi Anan
    - Кофи Анан, Bulgarian → Kofi Anan
    - كوفي عنان, Arabic → Kufi Anan
    - कोफी अन्नान, Hindi → Kofi Anan

Name normalisation – Why? (1)

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

  Владимир Устинов (Russian)
  - Vladimir Ustinov (English)
  - Wladimir Ustinow (German)
  - Vladimir Oustinov (French)

- Various ways to represent the same sound: sh, sch, ch, š, e.g.
  - Bašar al Assad
  - Sashar al Assad
  - Baschar al Assad
  - Bachar al Assad



Name normalisation – Why? (2)

- Diacritics are often omitted, e.g.
  - **Wałęsa** → 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

Name normalisation (3) - ~ 30 Rules

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

Malik al-Saïdoullaïev
Malik al-Saidoullaiev

Malik al-Saidoulaiev
Malik al-Saidulaiev
Malik Saidulaiev
… **mlk sdlv**

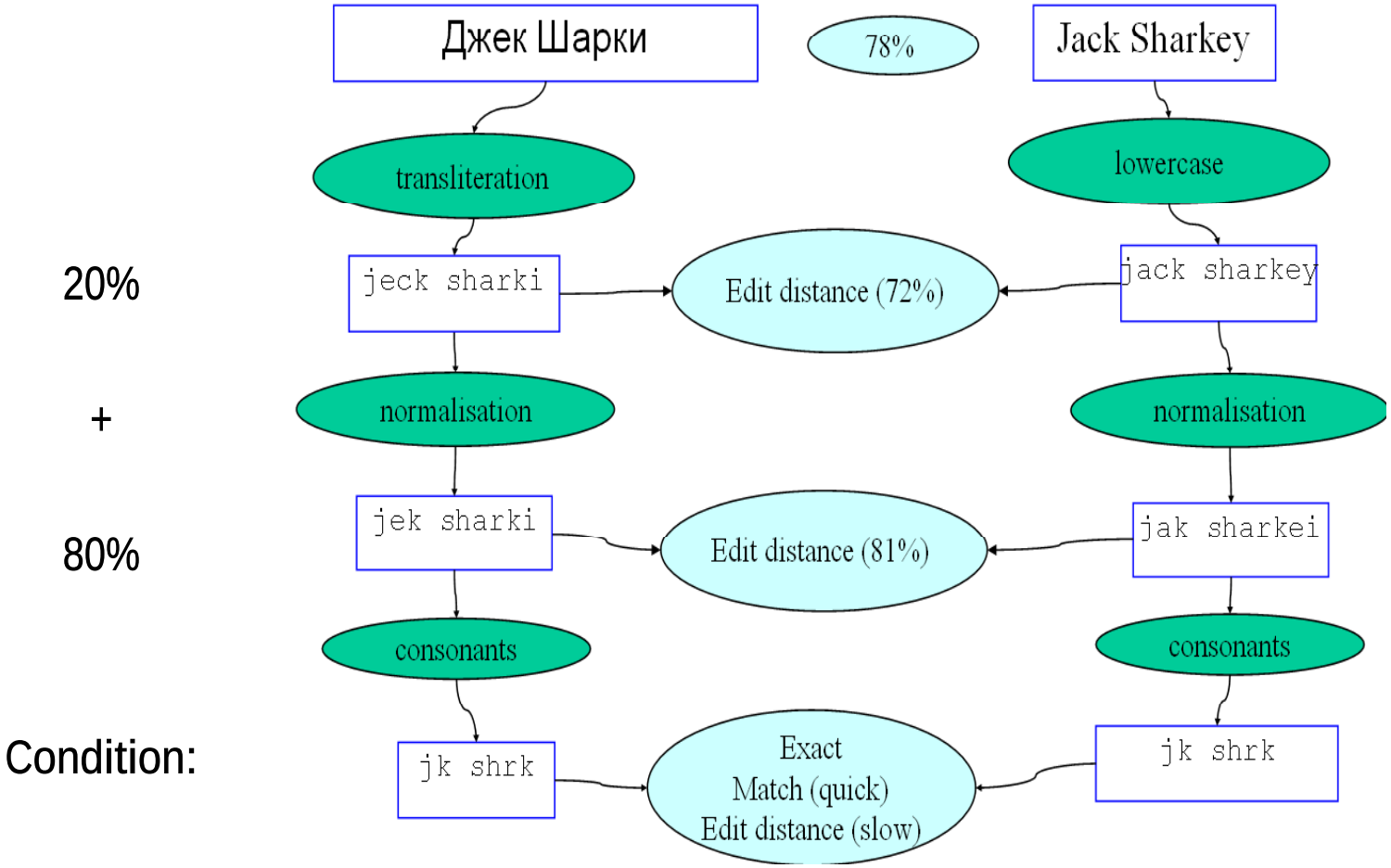
Name	Normalised form
Mohammed Siad Barre, Mohamed Siad Barré, Мохаммед Сиад Барре, محمد سياد بري	**mhmd sd br** (mohamed siad bare)
Mahmoud Ahmadinejad, Mahmūd Ahmadīnežād	**mhmd hmdnjd** (mahmud ahmadinejad)
Сергей Куприянов, Sergei Kupriyanov, Sergei Kuprianow, Sergueï Kouprianov	**srg kprnv** (sergei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

Similarity measure for name merging

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:



Merging name variants – some results

- 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

Person name recognition and variant merging – Result

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 Мушарраф (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

tomatically
dia

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)

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

Musharraf 't
cnn 21-NOV-(

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-0

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

[said]: The m
to improve the
instead,

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

Related Stories

Pakistan defends emergency rule
June 23, 2007 - November 22, 2007

Pakistan's president urges calm
May 5, 2007 - June 23, 2007

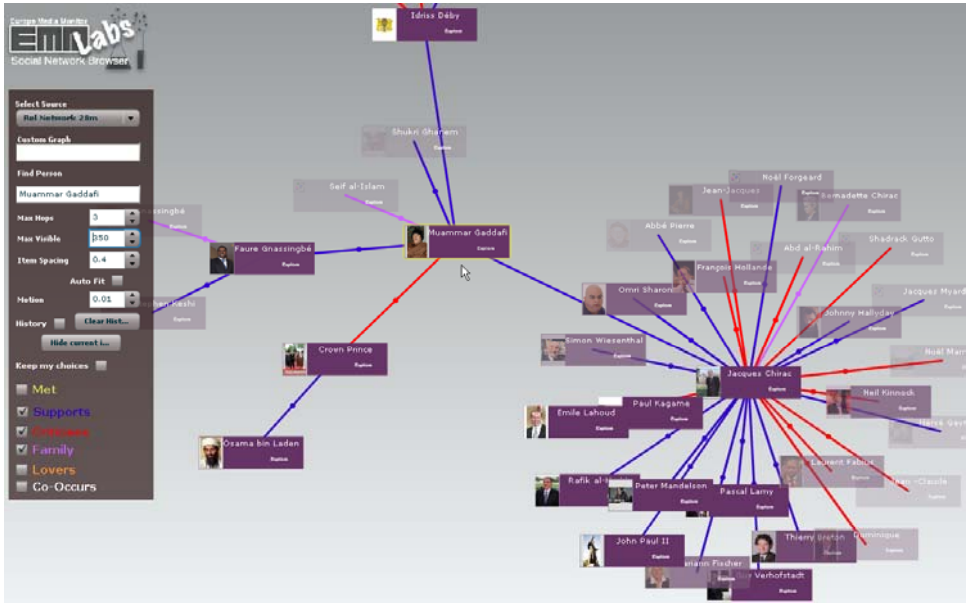
Troops launch hostage rescue bid
June 23, 2007 - October 25, 2007

Protests in Pakistan take aim at President Musharraf
March 10, 2007 - April 3, 2007

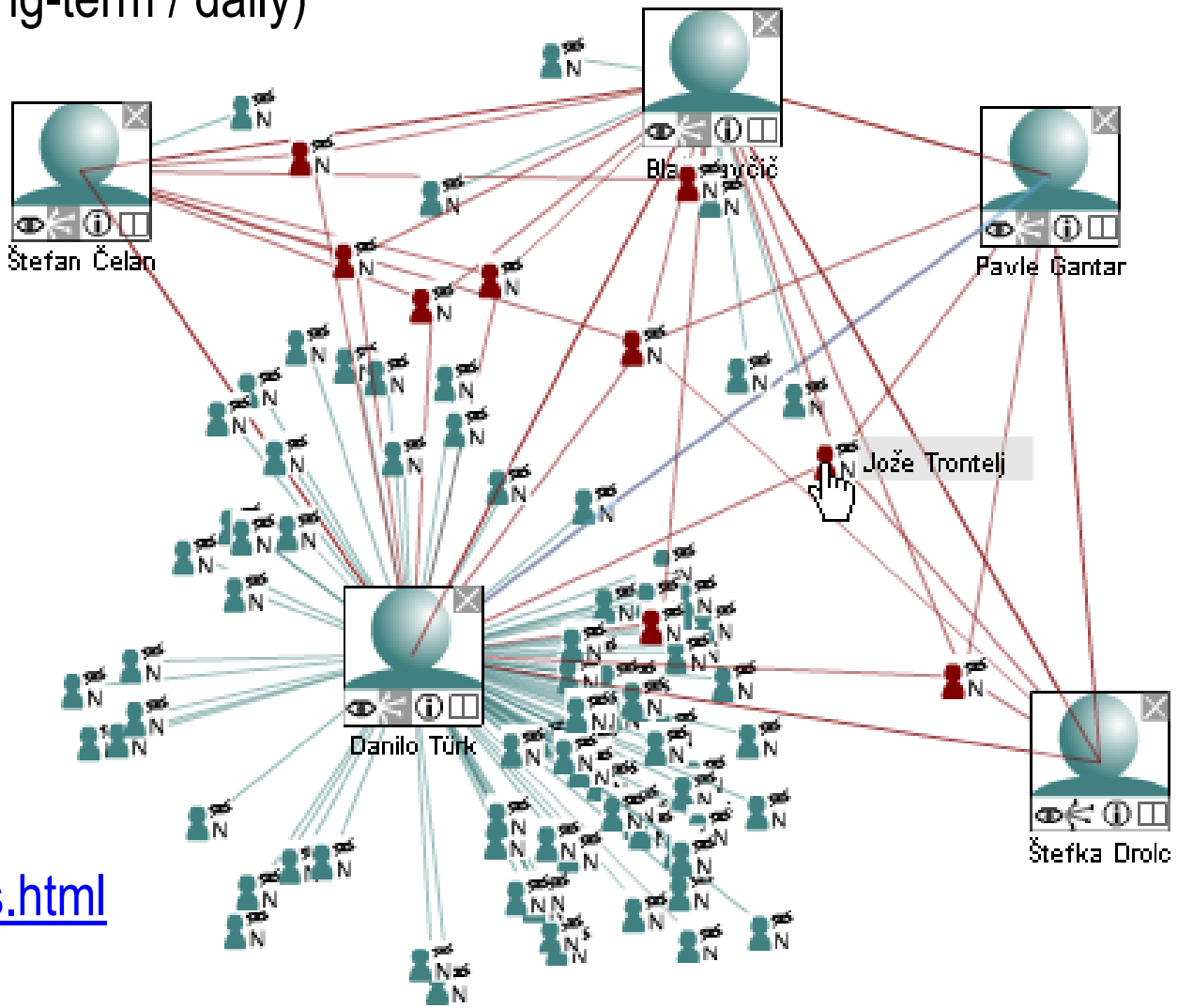
42 die in bomb attacks on 'lucky' weekend
July 15, 2007 - October 25, 2007

Demonstrations at UK embassy in Iran

Social network extraction from multilingual news



1. Co-occurrence networks (long-term / daily)



2. Quotation networks
   http://langtech.jrc.it/picNews.html
3. Relationship network
   http://emm-labs.jrc.it/LiveNews/Nets/Social.html

Associated names: weighting formula

$$w_{e_1,e_2} = \ln(C_{e_1,e_2}) \cdot \frac{2 \cdot C_{e_1,e_2}}{(C_{e_1} + C_{e_2})} \cdot \frac{1}{1 + \ln(A_{e_1} . A_{e_2})}$$

Number of clusters they appear in together

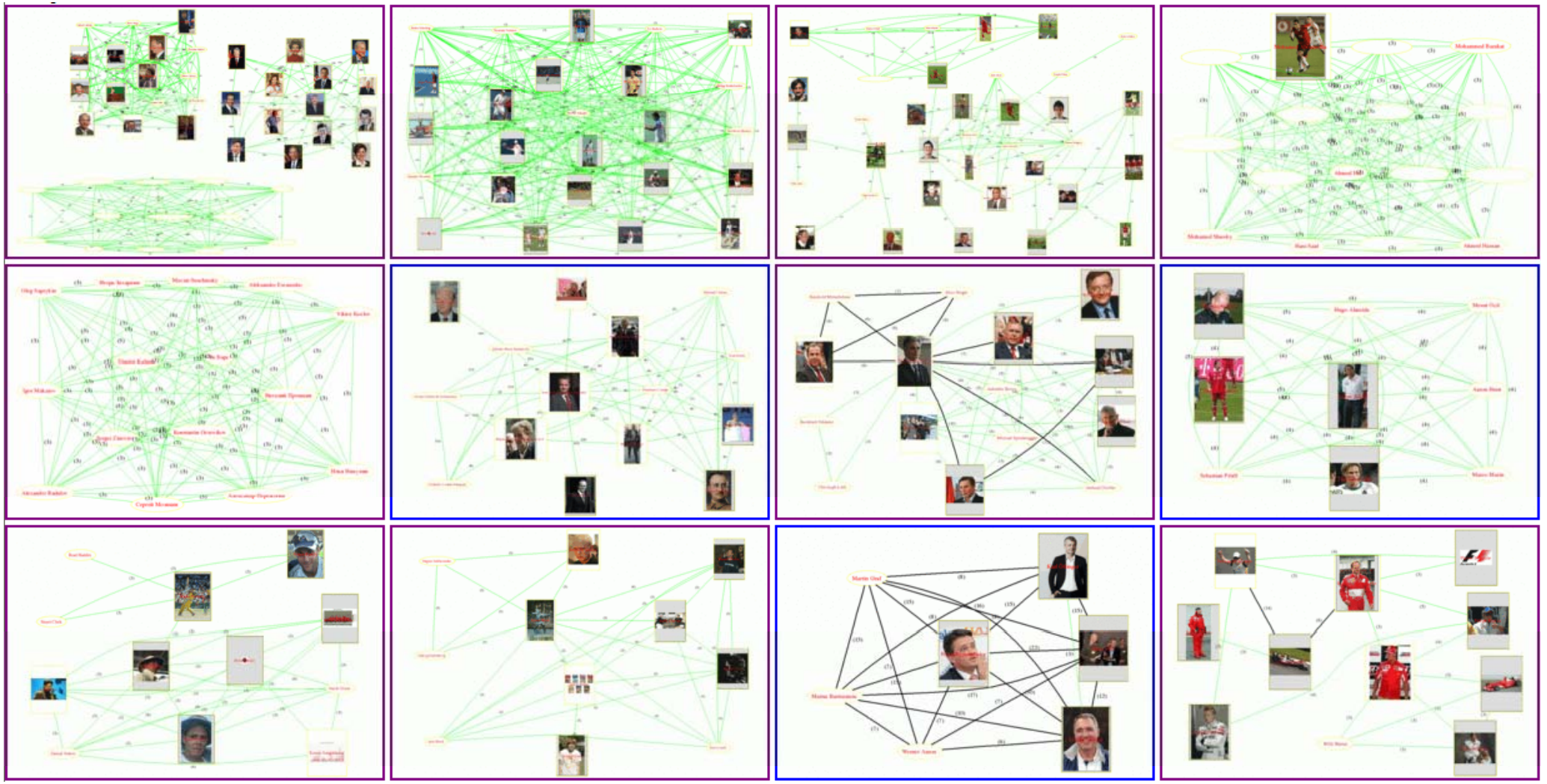
Weighting depending on the total number of clusters they appear in

Weighting depending on the total number of persons they are associated with

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)
Qazi Hussain Ahmed (0.9)

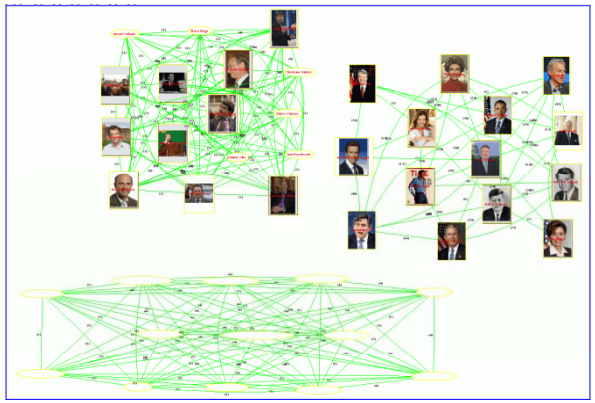
Daily co-occurrence networks (reset every midnight)

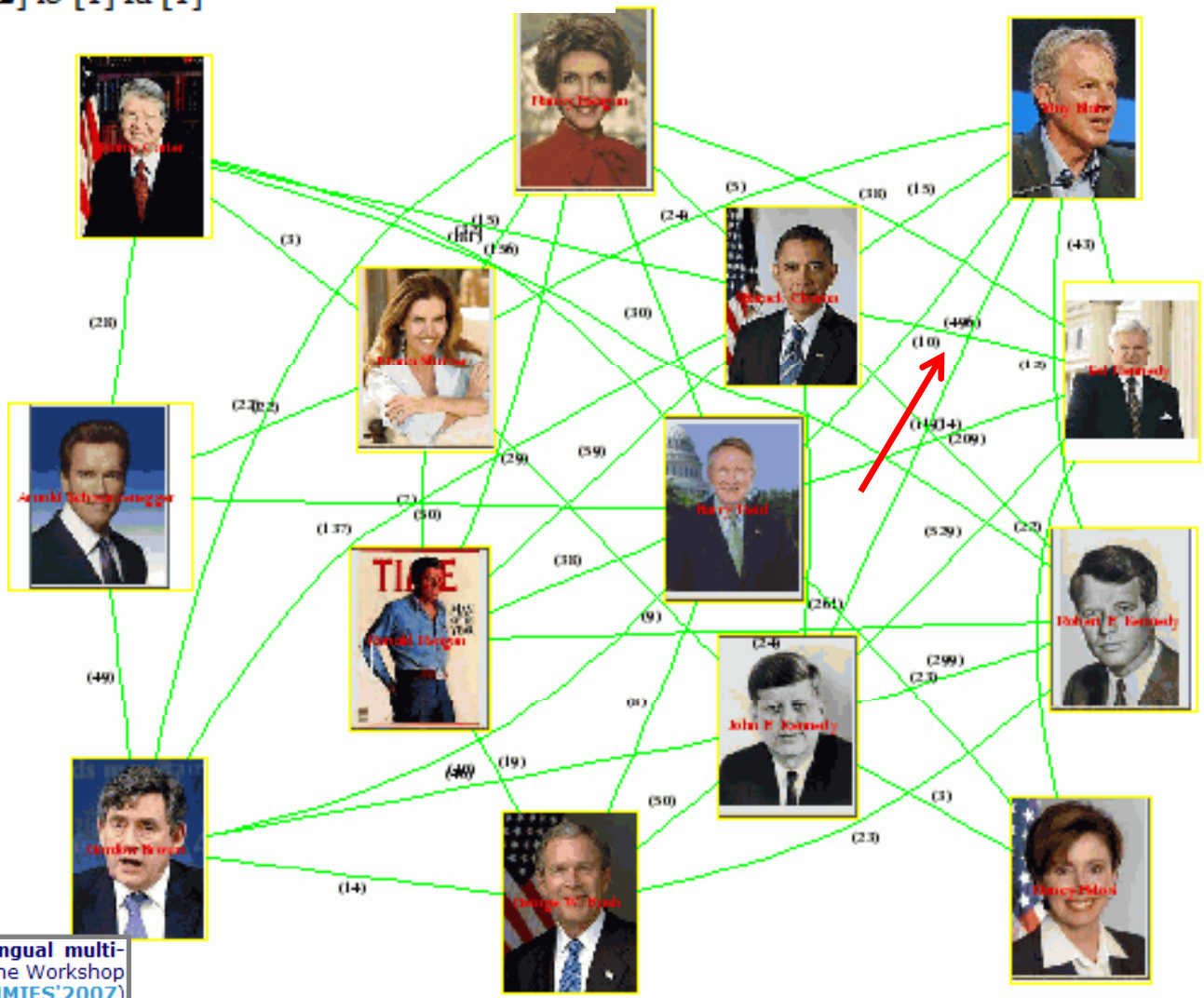


http://langtech.jrc.it/entities/socNet/last.html

Daily mixed-language co-occurrence networks (~100,000 links/day)

de [786] en [733] fr [592] ar [414] pt [225] sv [196] es [170] ru [145] nl [145] da [89]
no [81] sl [54] hu [53] it [52] ro [49] bg [37] fi [37] sk [16] ca [14] vi [12] sr [10] cs [9]
et [8] tr [5] lt [5] pl [4] mt [3] id [3] el [3] hr [2] lb [1] fa [1]





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)

Daily co-occurrence networks: Edward Kennedy – Barack Obama

499 articles in the news since midnight talking about those two persons.

Last updated **Wed, 26 Aug 2009 14:13:59 +0200**

Edward Kennedy ... Barack Obama

Rss for two Entities Wed, 26 Aug 2009 14:13:59 +0200 /var/www/cgi-bin/misc/rssTwoNames.pl V0.06 http://press.jrc.it/NewsExplorer/entities/en/226.html http://press.jrc.it/NewsExplorer/entities/en/1510.html Barack Obama Edward Kennedy

2009-08-26T14:11+0200 . Murió el senador Edward Kennedy **[clarin - es]**

... . El senador estadounidense [Edward M. Kennedy] ... , miembro de una de las familias más rele (...) dor de nuestro tiempo", dijo el presidente [Barack Obama] en un comunicado. El presidente se mostró & ...

2009-08-26T14:11+0200 . اوباما: وفاة كنيدي نهاية فصل هام في السياسة الامريكية **[bbc-arabic - ar]**

... عن حزنه الشديد لوفاة السيناتور الديمقراطي ! [باراك اوباما] أحرب الرئيس الامريكي ، احد أبرز رجالات السياسة الأمريكية، عن عمر السرطان. [ادوارد كنيدي] توفي السناتور الديمقراطي الامريكي ...

2009-08-26T14:10+0200 . National Review Online: Their Lion, Our Bane **[NPRnews - en]**

... the dead either. This is his remembrance of [Ted Kennedy] ... . There is a lot one could say (...) are reform and also the campaign of Senator [Barack Obama]. There are the personal failings and traged ...

2009-08-26T14:06+0200 . Ted Kennedy morre aos 77; conheça trajetória do irmão de John Kennedy **[folha - pt]**

... os 77 anos, o senador democrata americano [Edward M. Kennedy] ... , vítima de um câncer de cérebro, informo (...) ão democrata que confirmou a candidatura de [Barack Obama] à Presidência. Mesmo sofrendo de dores caus ...

2009-08-26T14:04+0200 . Senaattori Kennedy muistetaan liberaalina leijonana **[keskisuomalainen - fi]**

... u ympäri maailmaa. Yhdysvaltain presidentti [Barack Obama] kuvaili oloaan lohduttomaksi. – Viiden vuos (...) änellä diagnosoitiin aivosyöpä viime vuonna. [Ted Kennedy] ... valittiin senaattiin vuonna 1962 veljensä J ...

2009-08-26T14:01+0200 . Potere assoluto [La giornata] **[IlFoglio - it]**

... ni fa a Chappaquiddick la tragedia che segnò [Ted Kennedy] ... L'arroganza della stirpe più amata tolse al (...) ensi durante il pranzo di festeggiamento di [Barack Obama], dopo il giuramento presidenziale del 20 ge ...

2009-08-26T14:00+0200 . Ted Kennedy hylles av Obama **[hegnar - no]**

... bort tirsdag. Slik hylles han av psident Barack Obama. Senator [Edward Kennedy] ... , en av de mest sentrale og karismatiske se ...

2009-08-26T13:59+0200 . Senator Edward Kennedy dies at age 77 **[reuters - en]**

... BOSTON (Reuters) - U.S. Senator [Edward Kennedy] ... , a towering figure in the Democratic Party (...) Democrats as they seek to answer President [Barack Obama]'s call for an overhaul of the healthcare sy ...

2009-08-26T13:57+0200 . 09:00 ABD'nin en eski senatörü Edward Kennedy, öldü **[netgazete - tr]**

... tör ve Demokrat Parti'nin önemli şahsiyeti [Edward Kennedy] ... , 77 yaşında hayatını kaybetti. CNN, Amerik (...) nde de Demokrat Parti'nin aday adaylarından [Barack Obama]'yı desteklediğini açıkladı. Kennedy'nin des ...

2009-08-26T13:57+0200 . Edward „Ted" Kennedy - Der Tod einer Legende **[bildt-online - de]**

... 0 Senator [Edward Kennedy] ... , politisches Schwergewicht der USA und jün (...) tellt. Bei der Amtseinführung von Präsident [Barack Obama] (48) im vergangenen Januar brach Edward Ken ...

Quotation network, built from multilingual quotations

Chief Superintendent Stewart Gull said: "Wright, from Ipswich, has been charged with the murder of all five women."

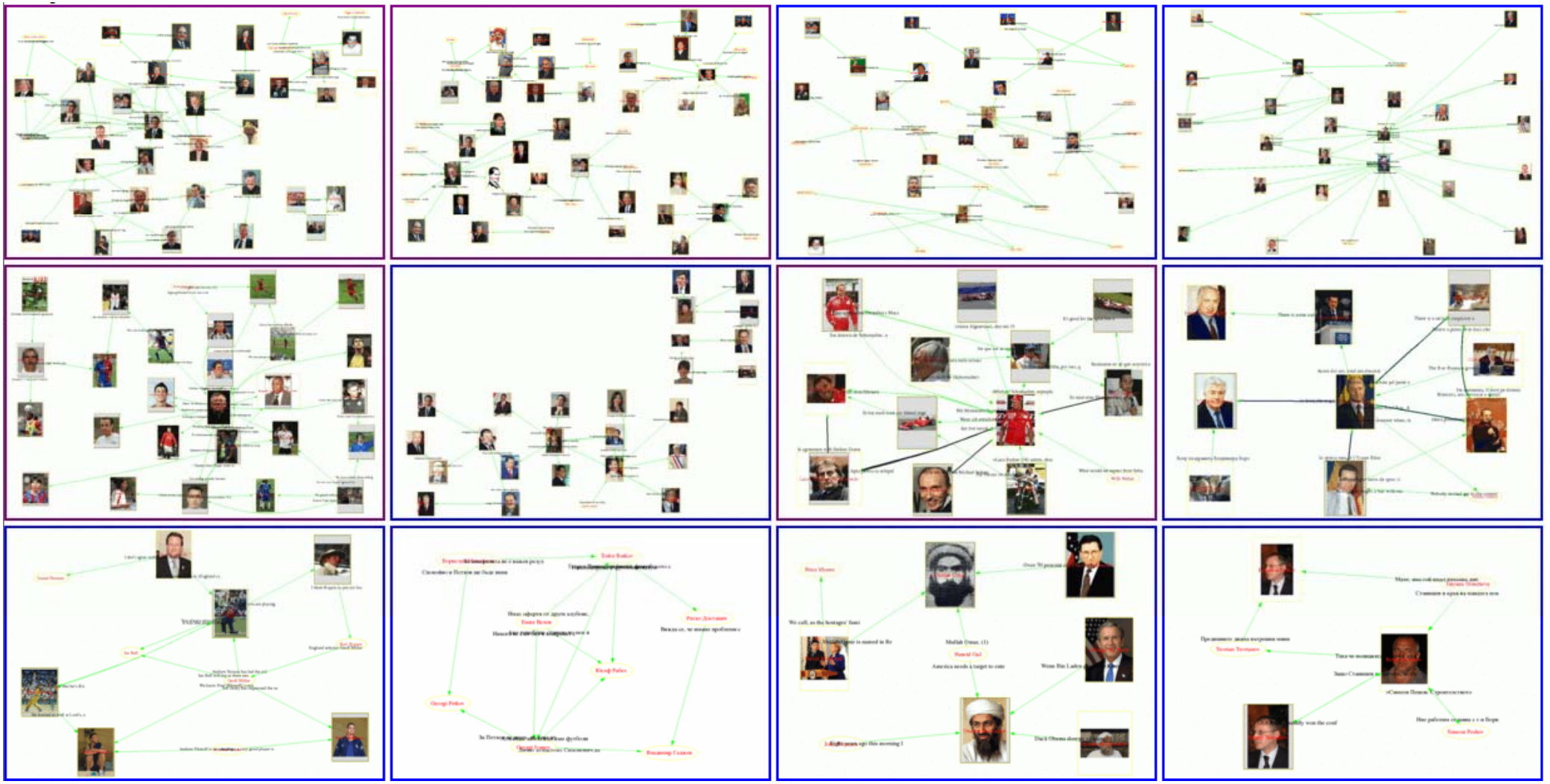
"I am against Turkey's full membership, and I have also told this to Turkish Prime Minister Erdogan," Sarkozy said, during a meeting with Angela Merkel.

Sarkozy, "Türkiye'nin AB'ye tam üye olmasına karşıyım. Bunu Başbakan Erdoğan'a da söyledim" dedi.

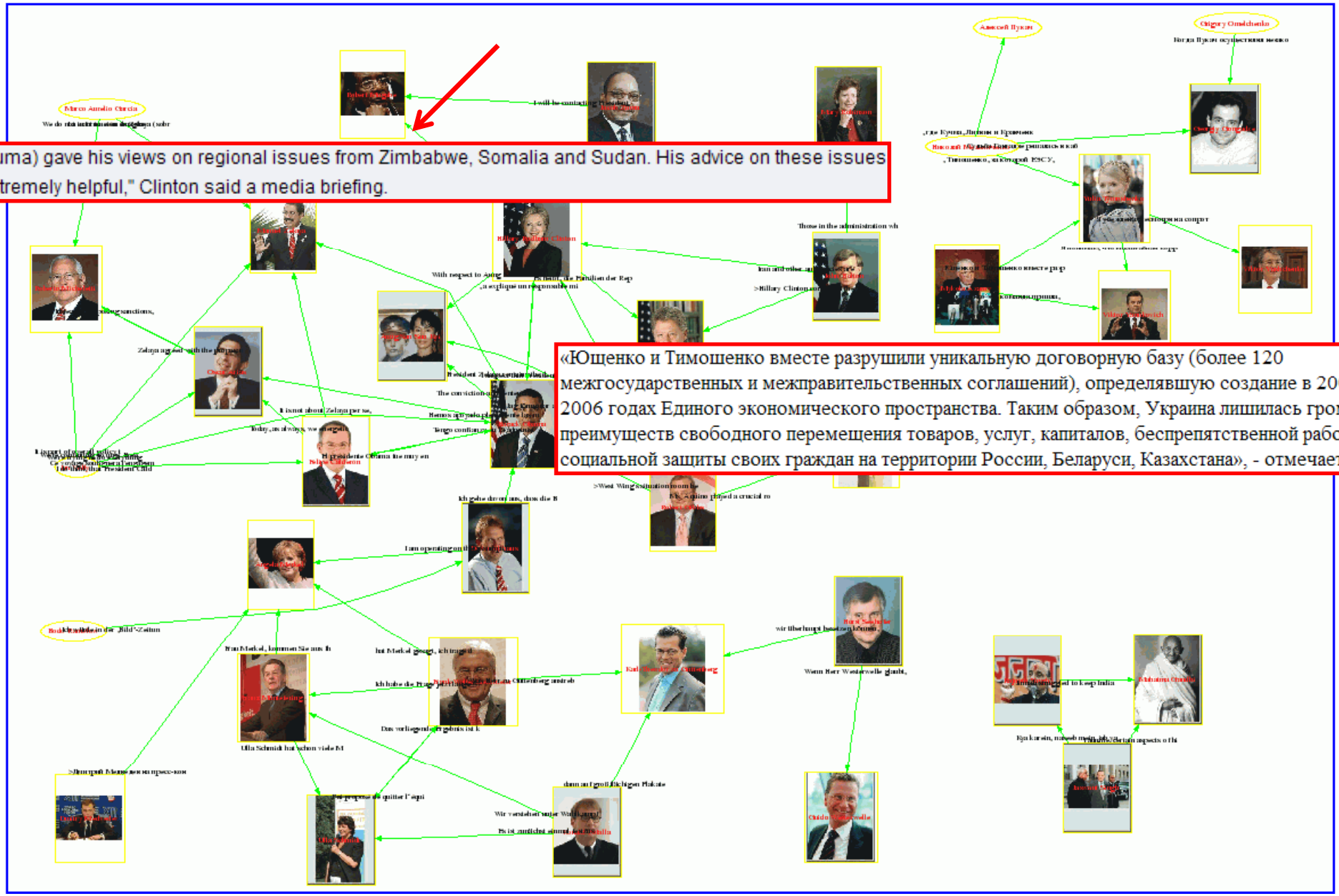
Gull → Wright (1)
Sarkozy → Erdogan (2)

Pouliquen Bruno, Ralf Steinberger & Clive Best (2007). **Automatic Detection of Quotations in Multilingual News**. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'2007), pp. 487-492. Borovets, Bulgaria, 27-29.09.2007. (PDF)

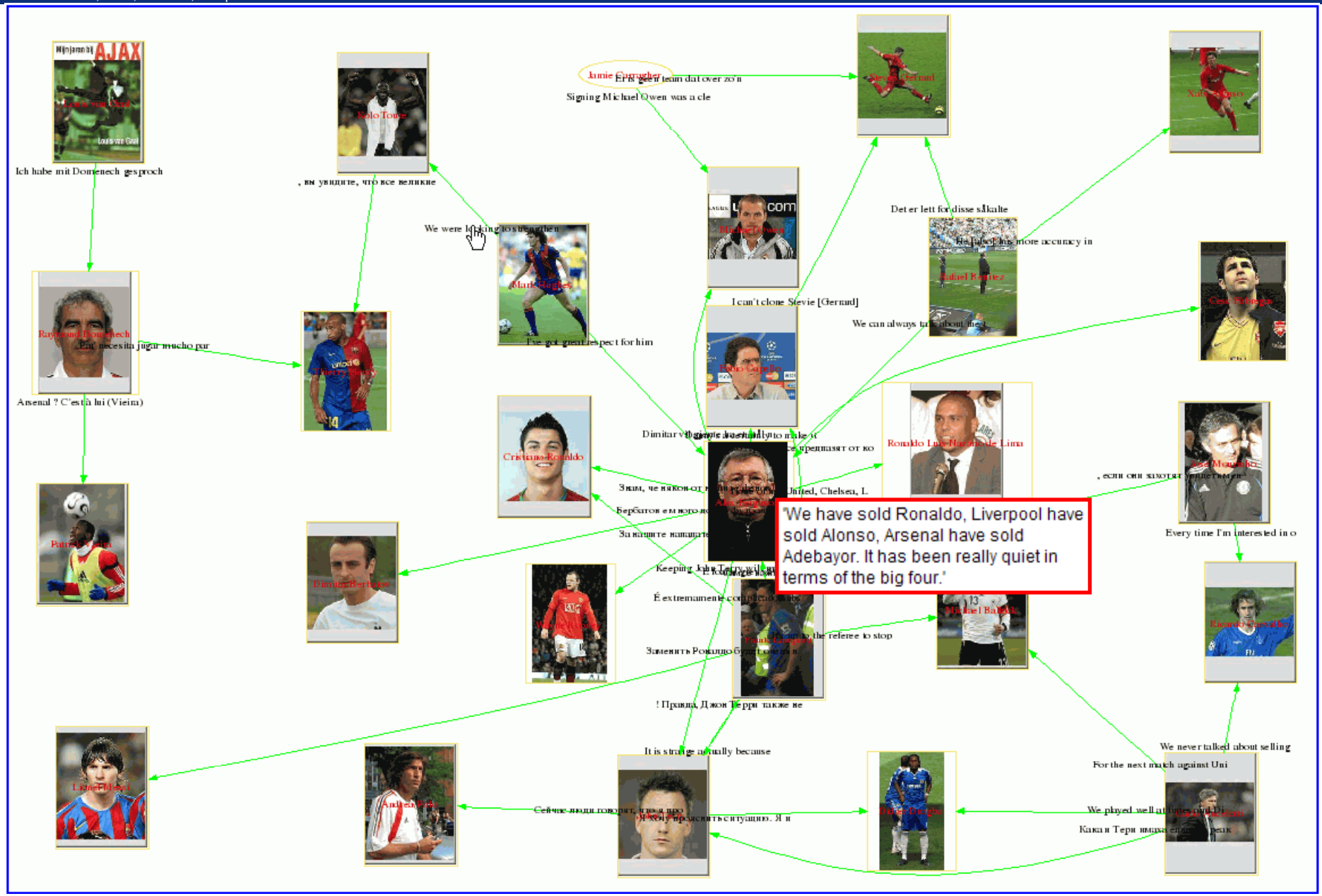
Quotation network: http://langtech.jrc.it/entities/socNet/quotes.html



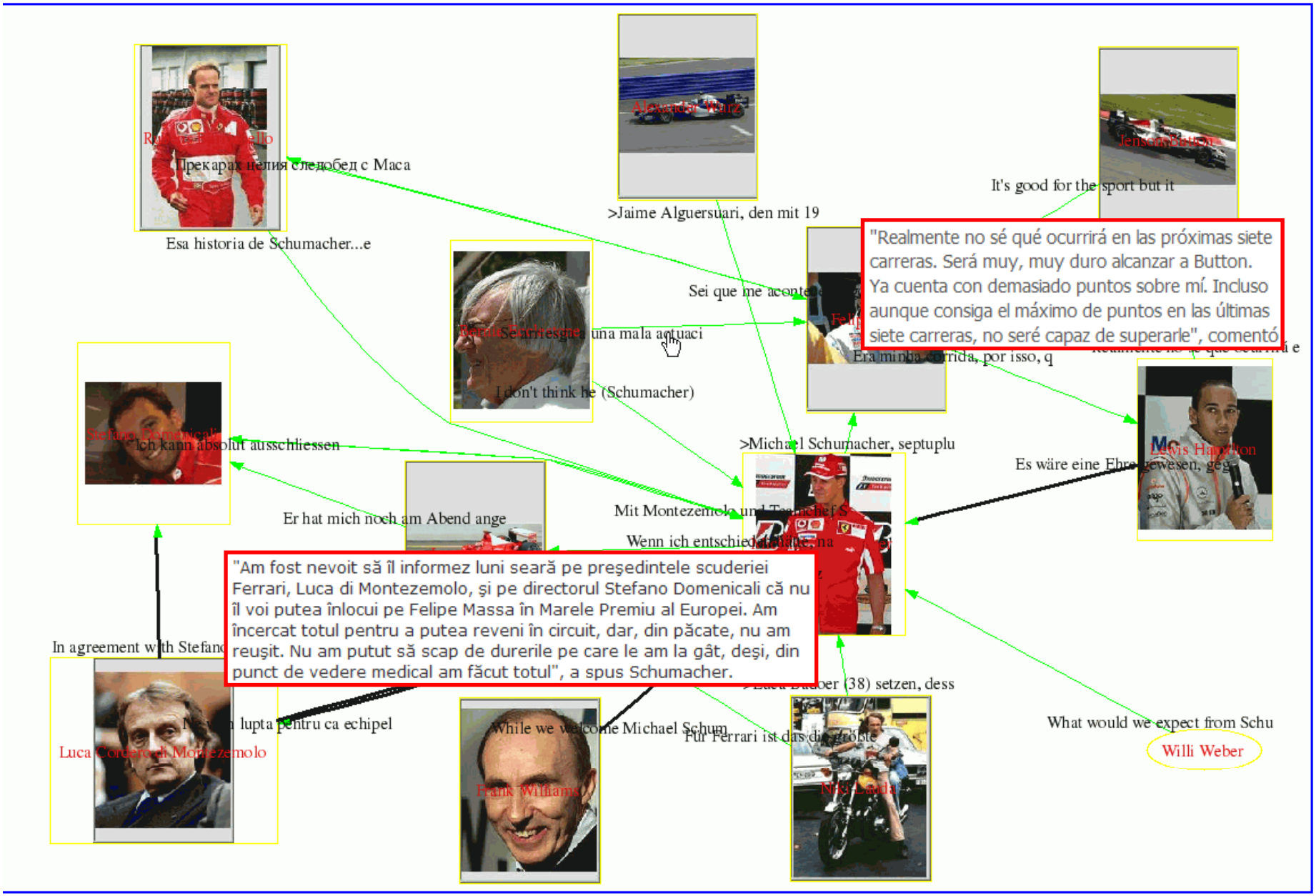
Multilingual quotation network (directed graph): sub-network



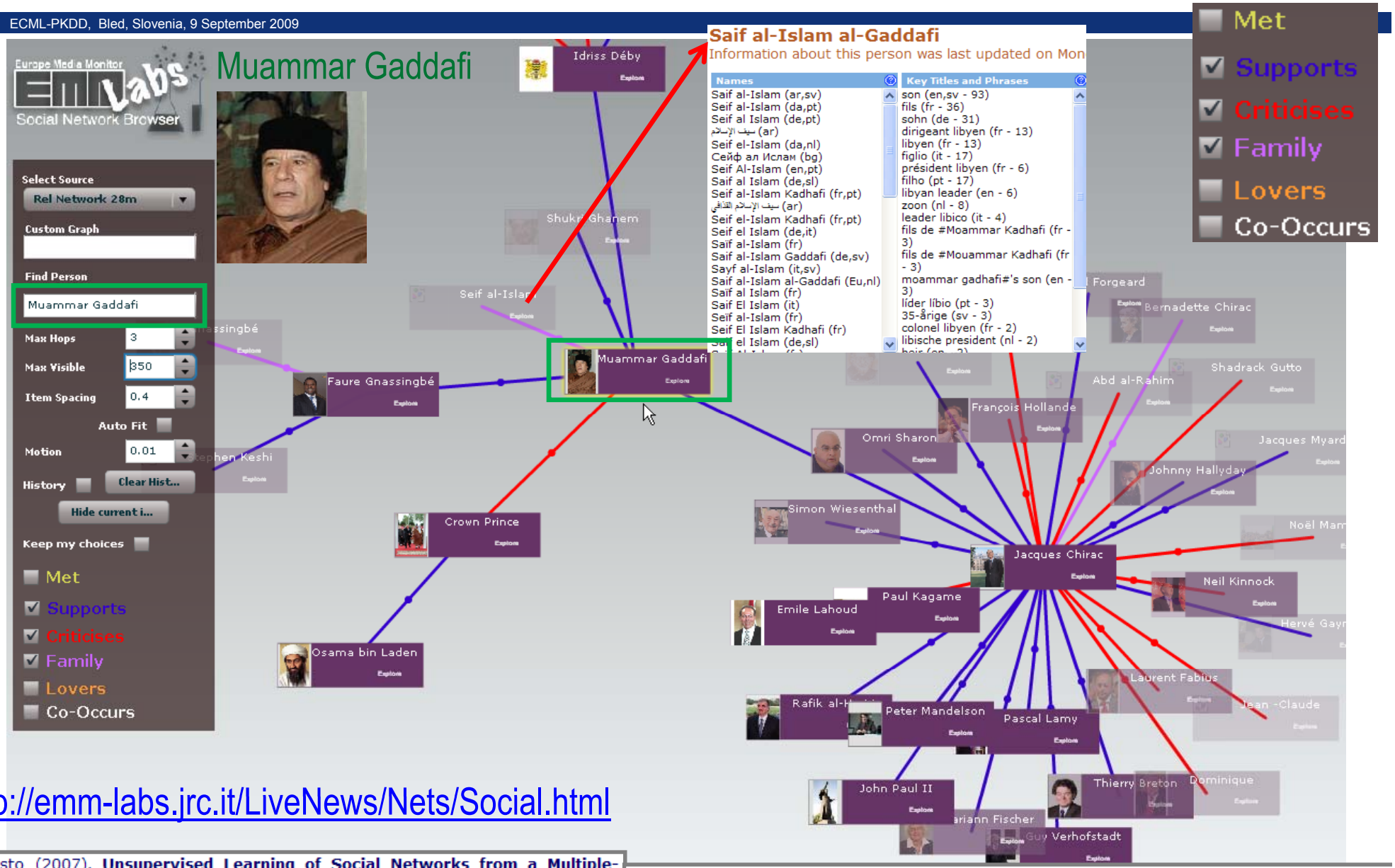
Quotation network: another sub-network



Quotation network: another sub-network



Network of specific relationships (English only)



http://emm-labs.jrc.it/LiveNews/Nets/Social.html

Tanev Hristo (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)

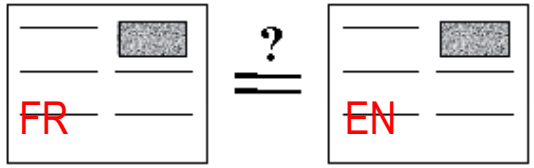
Pouliquen Bruno, Hristo Tanev & Martin Atkinson (2008). **Extracting and Learning Social Networks out of Multilingual News**. Proceedings of the social networks and application tools workshop (SocNet-08) pp. 13-16. Skalica, Slovakia, 19-21 September 2008. PDF.

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

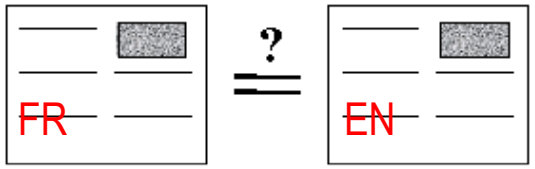
Fidel Castro announced his resignation as president of Cuba and commander-in-chief of Cuba's military on Tuesday, according to a letter published by state-run newspaper Granma.
cnn 9:23:00 AM CET

Approaches to cross-lingual document similarity calculation (1)



- 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

Approaches to cross-lingual document similarity calculation (2)



- 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)

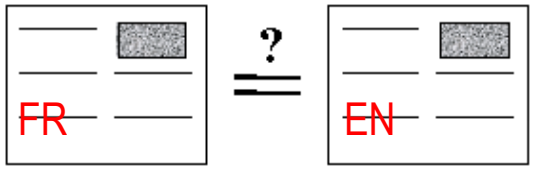
\+ Achieved results are relatively good

\- Bilingual approach is restricted to a few languages (OK for English as target lang.):

Language pairs = N * (N -1) / 2 (N = number of languages)

- NewsExplorer: 19 languages → 171 language pairs (342 language pair directions)!

Approaches to cross-lingual document similarity calculation (3)



- Alternative: use entities and thesauri as anchors:
  - Names of persons and organisations
  - Names of locations
  - Nodes from a multilingual thesaurus (e.g. Eurovoc)
  - 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.

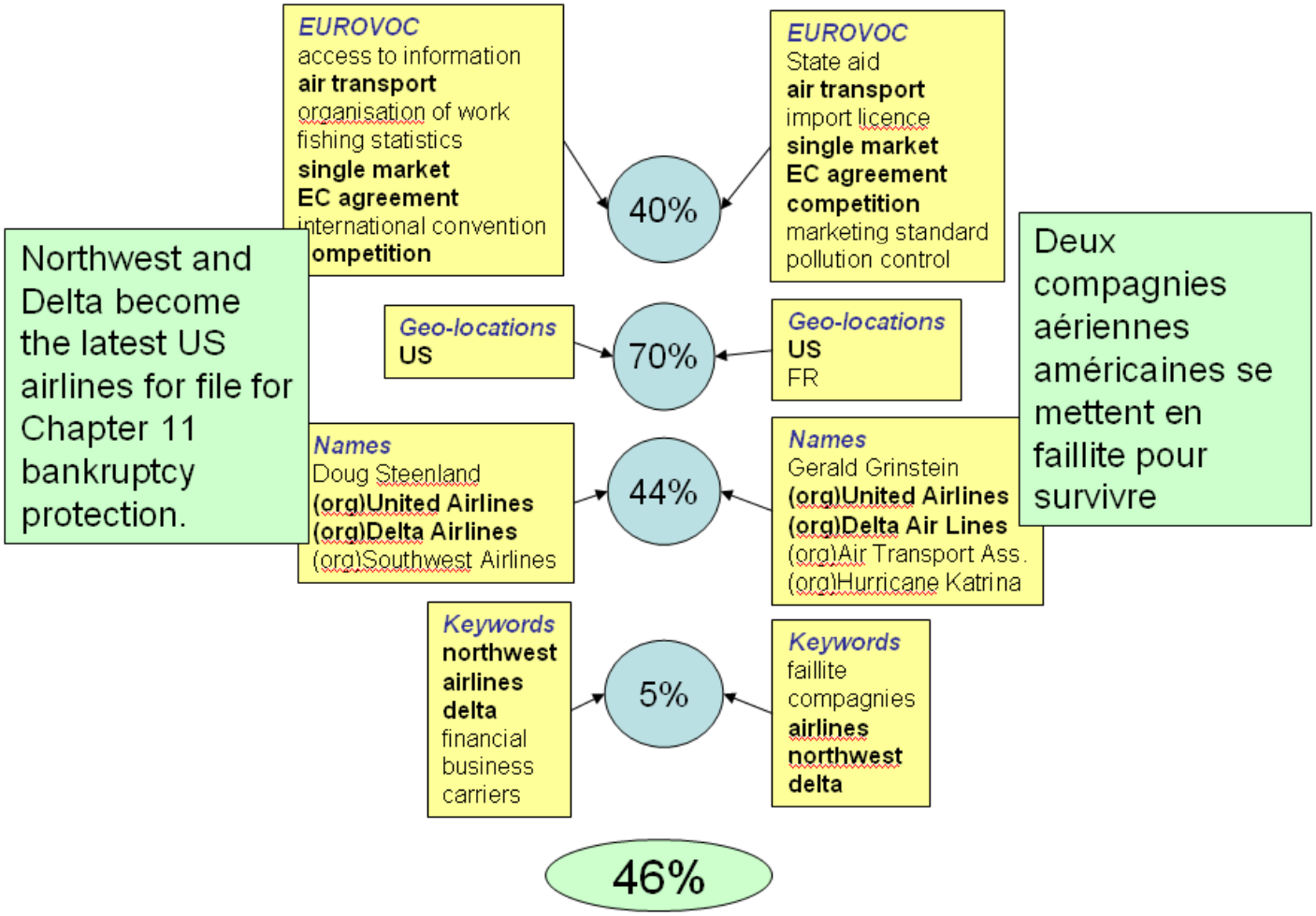
NewsExplorer - Cross-lingual cluster linking

Language-independent features for multilingual document representation

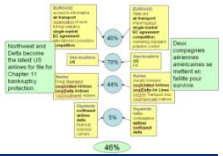
No MT or bilingual dictionaries

19 languages

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



Multilingual Geo-tagging



- 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	*Tony:* USA
	Blair: Malawi
Kofi Annan	*Kofi:* Mali
	Annan: Scotland
Javier Solana	*Javier:* Spain
	Solana: Philippines

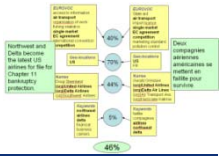
Pouliquen Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best (2006). **Geocoding multilingual texts: Recognition, Disambiguation and Visualisation**. Proceedings of the 5th International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

Eurovoc subject domain categorisation

eurovoc THESAURUS

Descriptor ID	Descriptor text	Inverse square Sum Tfidf^2	Cosine	Rank Cosine	Okapi	Rank Okapi	Rank	Prec	Rec
5641040706000000	FISHING CONTROLS [G]	.000144033	0.360	1	95.169	1	1	100	10
5641020000000000	FISHING GROUNDS [nt]	.00243484	0.308	2	65.018	14	2	100	20
5641040200000000	COMMON FISHERIES POLICY [G]	.00018023	0.280	3	62.910	20	3	100	30
5641040100000000	FISHERY MANAGEMENT [nt]	.000207080	0.279	4	79.302	6	4	100	40
5641040700000000	FISHING REGULATIONS [G]	.000197934	0.270	5	79.982	5	5	100	50
5641040704000000	FISHING PERMIT [G]	.00306631	0.261	6	71.577	8	6	100	60
5641040101000000	CONSERVATION OF FISH STOCKS [S]	.000189818	0.253	7	83.982	3	7	85	60
5641040600000000	FISHING AREA [G]	.000182474	0.252	8	84.178	2	8	75	60
5206040100000000	CONSERVATION OF RESOURCES [S]	.000234209	0.251	9	55.311	26	9	66	60
5641050000000000	FISHERY RESOURCES	.000402883	0.232	10	75.046	7	10	60	60
5641040000000000	CATCH OF FISH	.000313101	0.213	11	67.687	9	11	54	60
5641040000000000	FISHERIES POLICY	.00258399	0.203	12	58.416	23	12	50	60
5641040705000000	FISHING LICENCE	.000371136	0.181	13	57.618	25	13	46	60
5641060100000000	FISHING FLEET	.00106478	0.179	14	63.323	19	14	42	60
5641010000000000	FISHING INDUSTRY	.000551953	0.176	15	39.228	43	15	40	60
5641040201000000	EUROPECHE	.000738822	0.176	16	62.240	21	16	37	60
...									

Multi-label indexing with Eurovoc



- 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).

Assignment Result (Example)

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

Two freely available multilingual resources (22 languages)

- DGT-TM translation memory
  - For all 231 language pairs (462 language pair directions)
  - Made available by DGT, through JRC's website.

- JRC-Acquis multilingual parallel corpus (v 3.0)
  - Made available by JRC
  - Sentence-aligned for all 231 language pairs using two different aligners
  - Over 1 Billion words
  - Manually subject domain-classified using the Eurovoc thesaurus.

- Freely available for research purposes on our web site:
  http://langtech.jrc.it/ (go to 'Resources')

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006). **The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.** Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 2142-2147. Genoa, Italy, 24-26 May 2006.

NewsExplorer - Cross-lingual cluster linking

Language-independent features for multilingual document representation

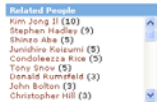
No MT or bilingual dictionaries

19 languages

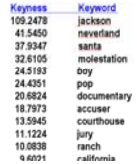
$$CLDS = \alpha \cdot S1 + \beta \cdot S2 + \gamma \cdot S3 + \delta \cdot S4$$

Sim1 (40%):
Multilingual Eurovoc subject domains

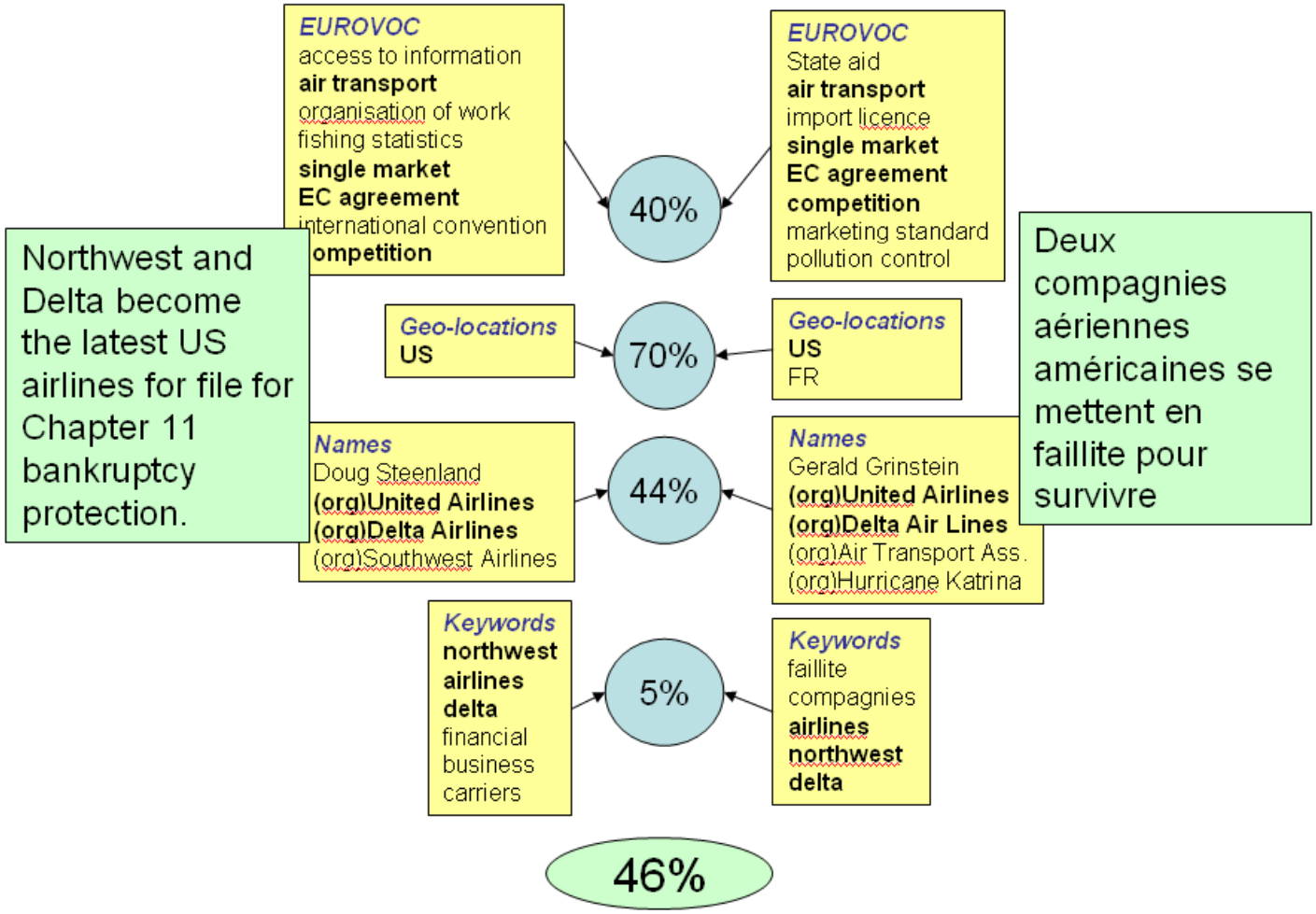
Sim2 (30%):
Geo-locations



Sim3 (20%):
Names + variants



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



Agenda

- Media Monitoring and multilinguality
- Europe Media Monitor (EMM) applications - Functionality
  - Publicly accessible at http://press.jrc.it/overview.html

- Some language technology components in detail
  - Multilingual person name recognition
  - Name variant matching across many languages
  - Social networks based on multilingual information extraction
  - Cross-lingual cluster linking, incl. multi-label categorisation using Eurovoc
- Conclusion and Ongoing work

Conclusion
Ongoing work

Summary

- Europe Media Monitor (EMM) applications - Functionality

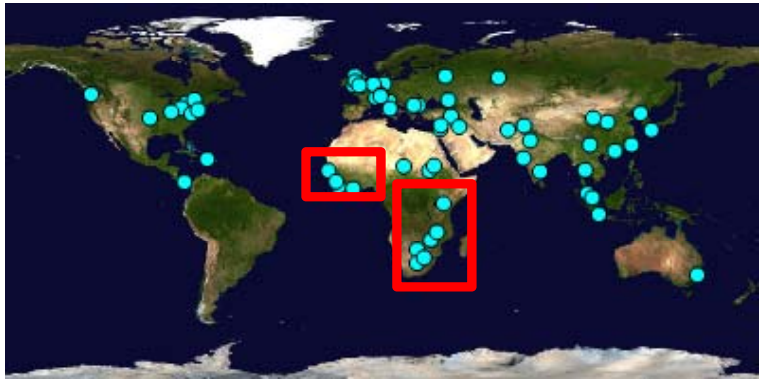
- Some language technology components in detail
  - Multilingual person name recognition
  - Name variant matching across many languages
  - Social networks based on multilingual information extraction
  - Cross-lingual cluster linking

→ Importance of multilinguality

Multilinguality: complementary news coverage in various languages

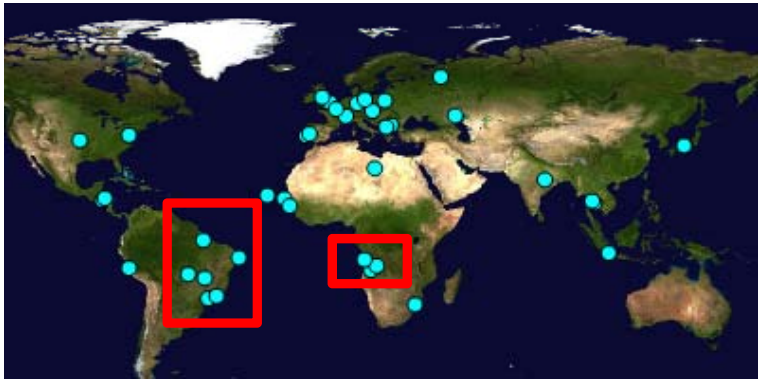
Locations mentioned in MedISys medical articles across languages – complementary coverage

Italian - German



English - French

Spanish - Portuguese



Multilinguality: More information about relations between people

Social networks 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)
- 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)
- Qazi 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)

Multilinguality: Gathering *more information* about people

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) Alexander Litwinenko (de) Alexandre Litvinenko (fr) Aleksandr Litvinenko (fi,no) Aleksander Litvinenko (nl,sv) Александра Литвиненко (ru) Александр Литвиненко (ru) Alexander Livlinenko (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)	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)	 *Image obtained automatically from Wikipedia* Read Wikipedia entry

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.

Ongoing work

- Ongoing activities:
  - Further integration of the four applications
  - Blog monitoring
- Current research:
  - Sentiment analysis for the news
  - Multilingual multi-document summarisation
  - Challenge: Using simple methods that do not require many language-specific resources