EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Language Technology

Linking News Content Across Languages

NODALIDA 2009

Odense, Denmark, 16 May 2009

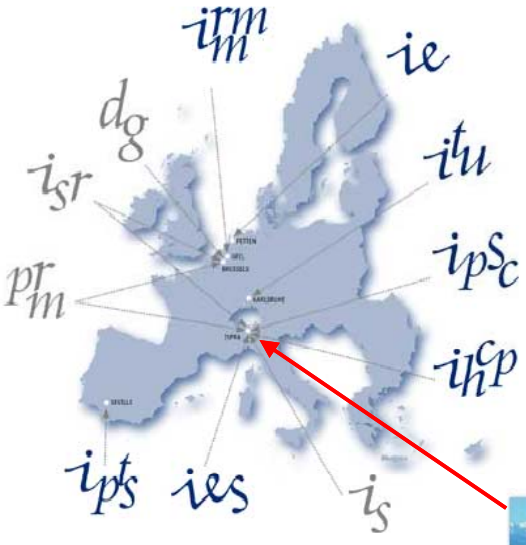
Ralf Steinberger

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

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

JRC
EUROPEAN COMMISSION

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

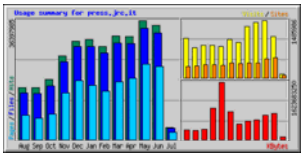
- Europe Media Monitor (EMM) applications
  - Publicly accessible at http://press.jrc.it/overview.html

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

Europe Media Monitor (EMM) family of applications

- 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

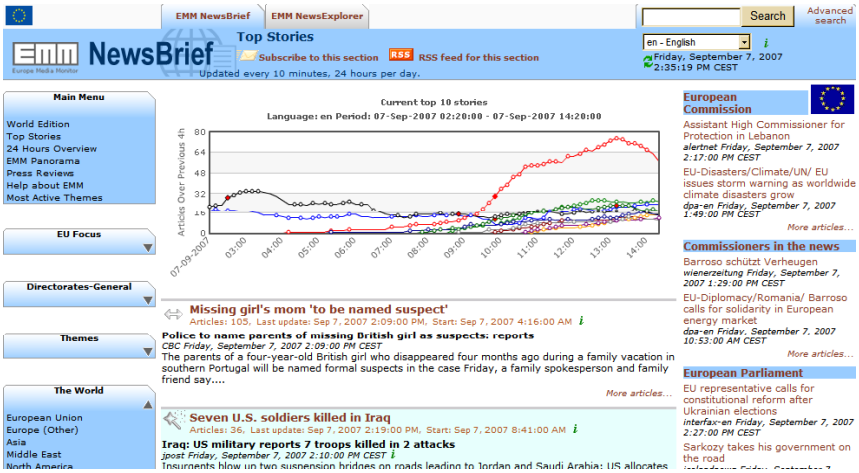
  - Combined between 1 and 2 Million hits per day
  - 30,000 – 50,000 distinct users per day



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, 20-124 April 2009. (PDF)

EMM – NewsBrief (up to 43 languages)

- 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-monolingually)
- Sends out email notifications for each category
- Detects breaking news
- Short-term story tracking

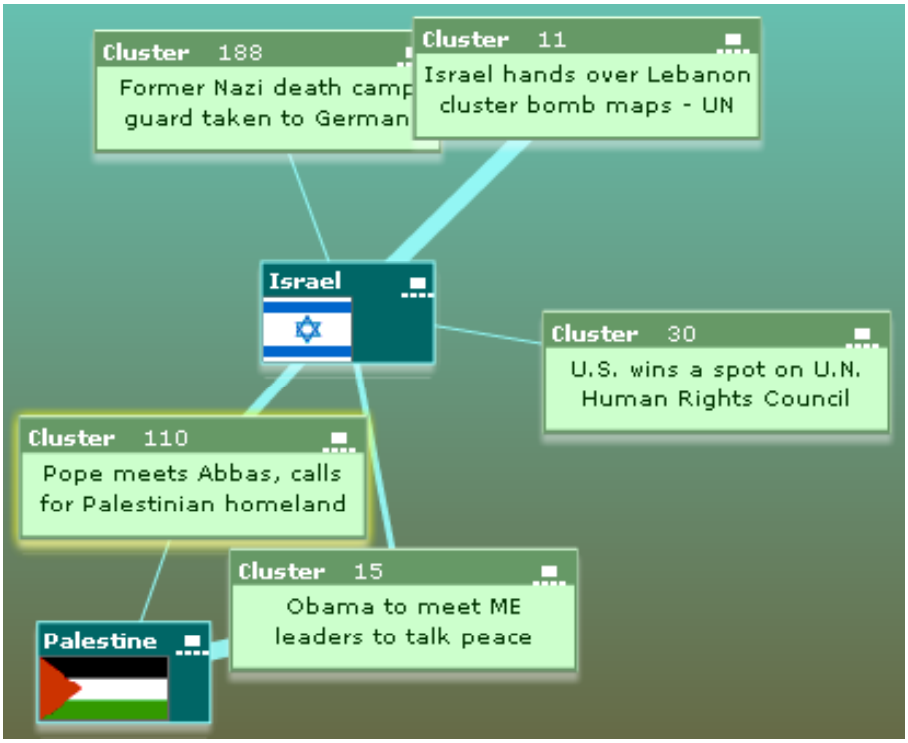


Atkinson Martin & Erik Van der Goot (2009). **Near Real Time Information Mining in Multilingual News**. Proceedings of the 18th International World Wide Web Conference (WWW'2009), pp. 1153-1154. Madrid, 20-124 April 2009. (PDF)

JRC EUROPEAN COMMISSION
NODALIDA-2009, Odense, Denmark, 16 May 2009

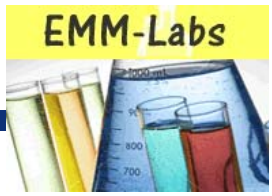
EMM-Labs: Real-time Cluster Network

Display of relations between clusters and categories

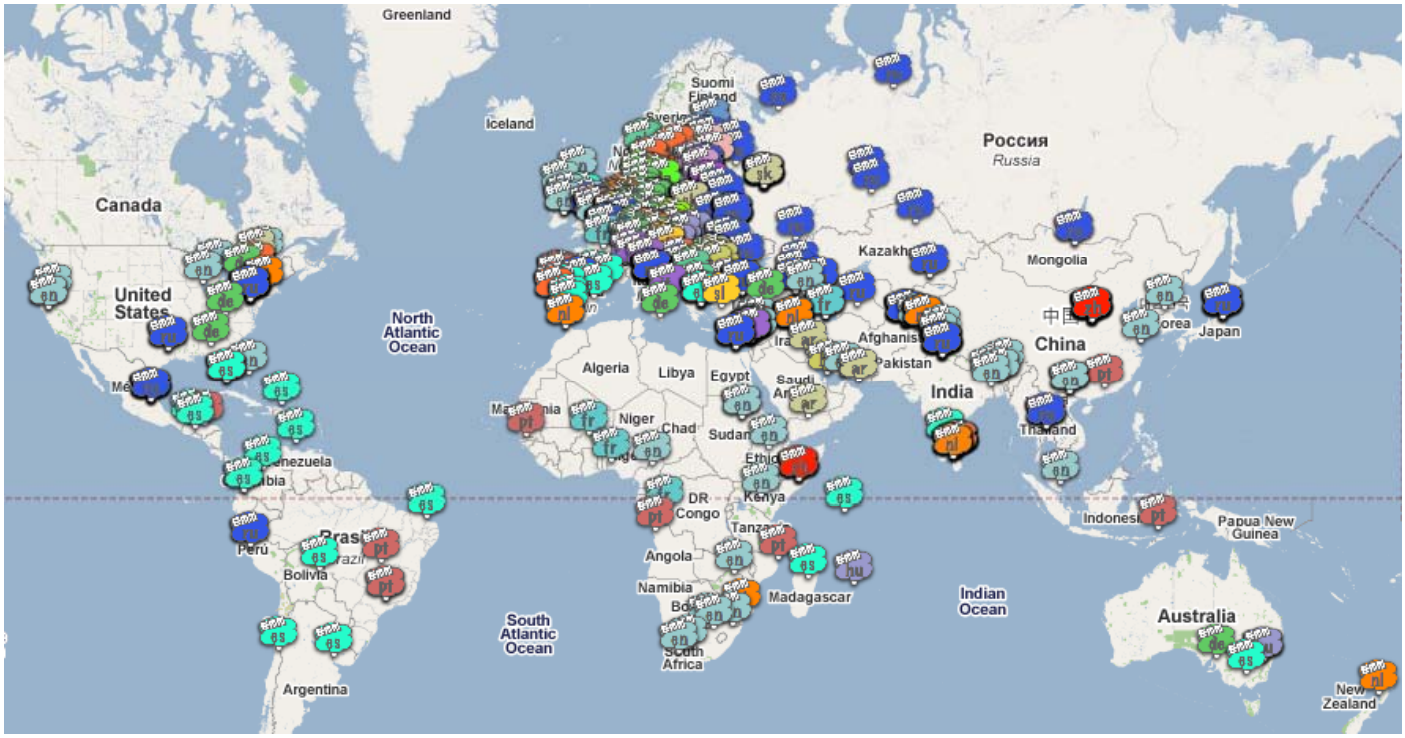


JRC EUROPEAN COMMISSION

EMM-Labs: NewsBrief Live Cluster Map

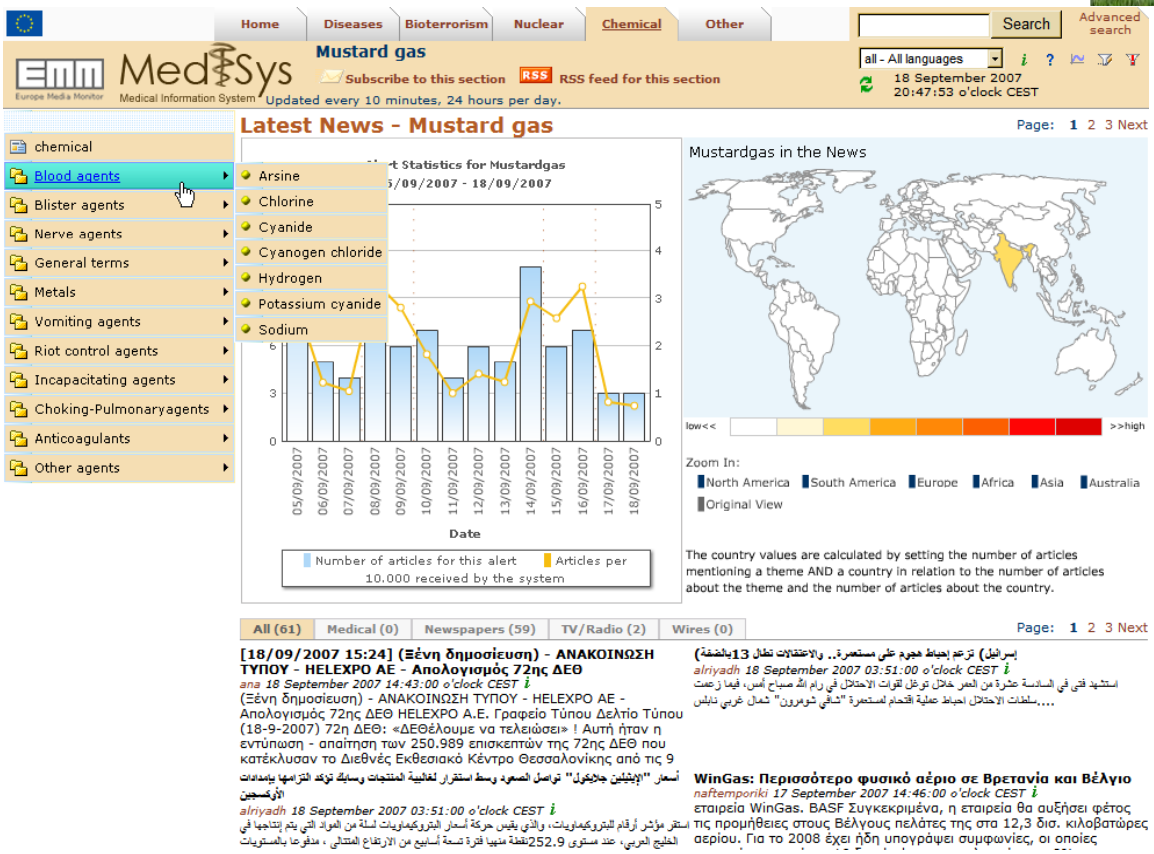


Display of latest geo-located news clusters



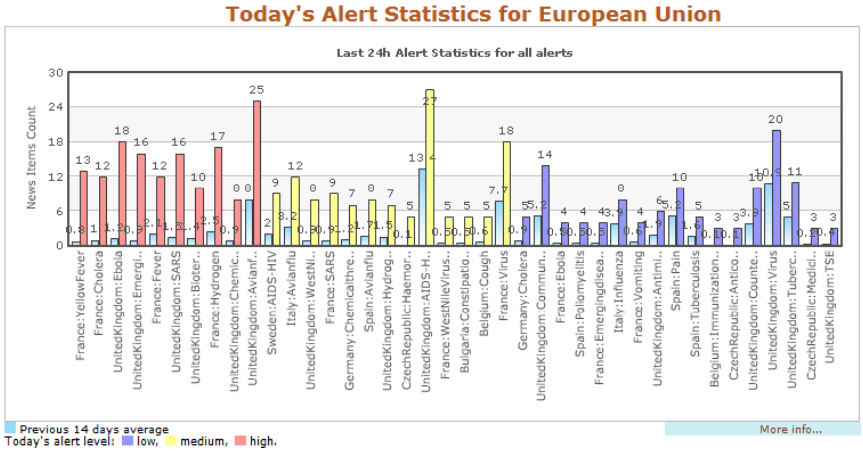
JRC EUROPEAN COMMISSION

MedISys – Filtering and classification in up to 43 languages



MedISys - Aggregation of multilingual information

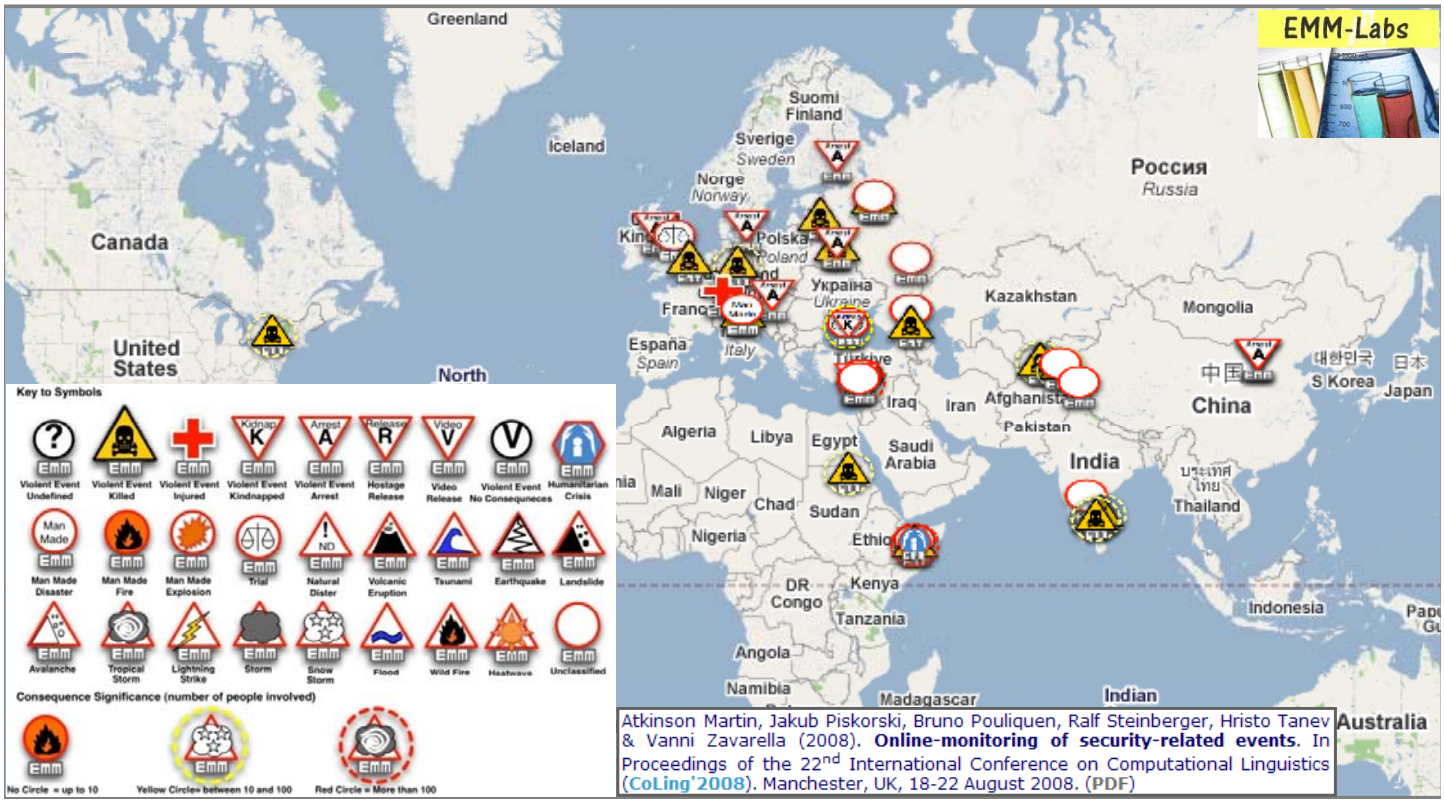
- Documents from all languages get classified according to the same countries and categories.
- An increase of the number of media reports on any country-category combination is detected,
- independently of the reporting language.
- Graphs and alerts may show events not yet reported in your own language.



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

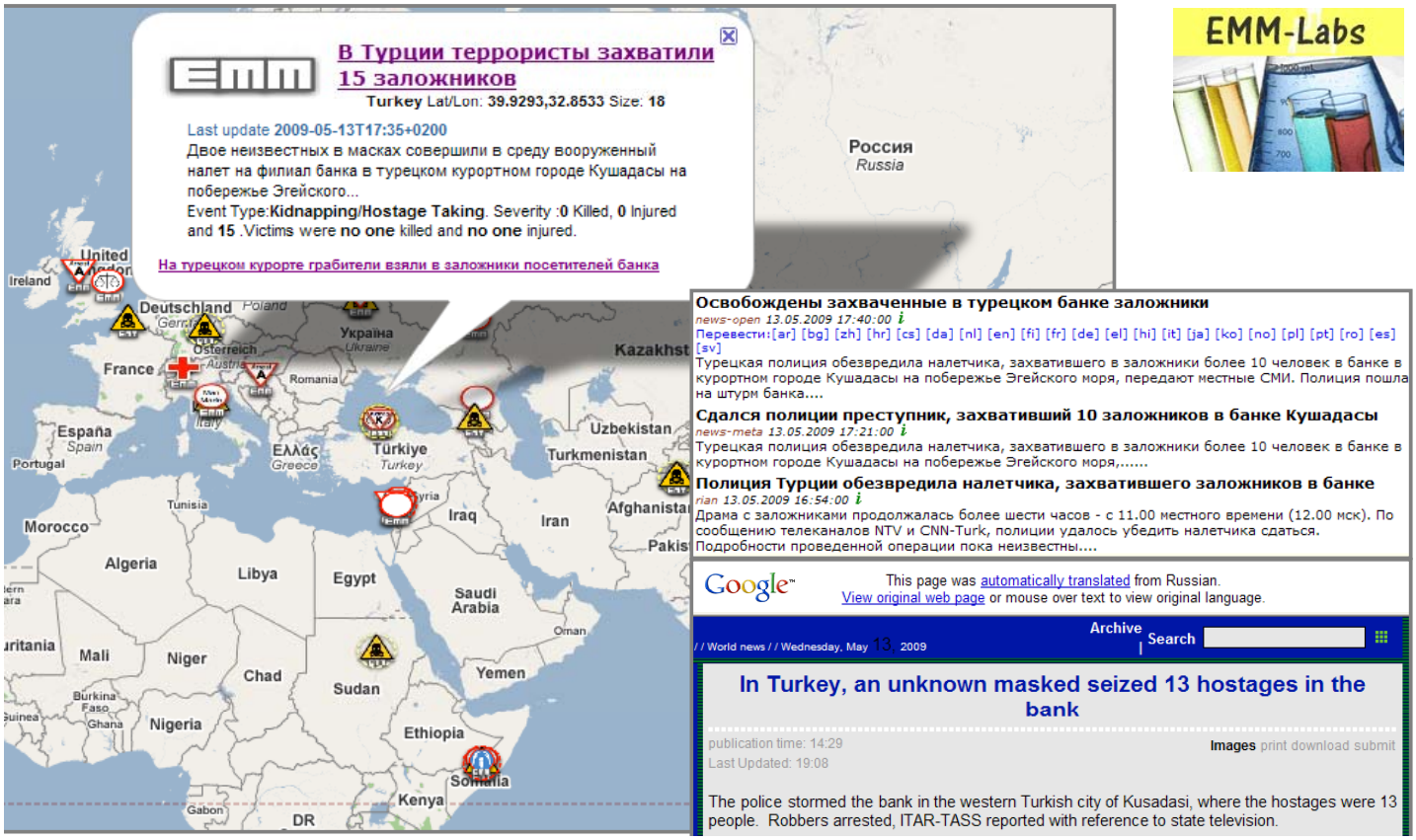
Detection and visualisation of events (violence/disasters/humanitarian/…)

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



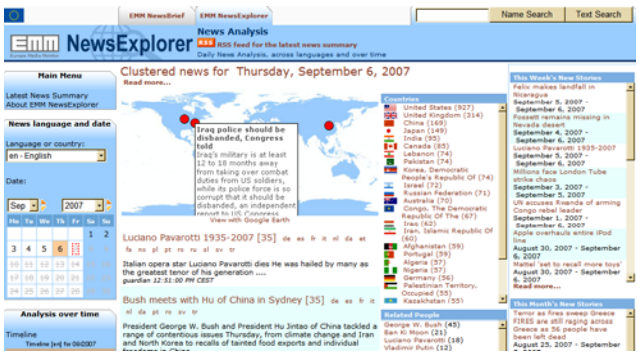
Atkinson Martin, Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger, Hristo Tanev & Vanni Zavarella (2008). **Online-monitoring of security-related events**. In Proceedings of the 22nd International Conference on Computational Linguistics (CoLing'2008). Manchester, UK, 18-22 August 2008. (PDF)

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



NewsExplorer – Functionality

http://press.jrc.it/NewsExplorer



NewsExplorer – Multilingual daily news overview

NewsExplorer – Cross-lingual cluster linking

NewsExplorer

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

Fidel Castro announced his resignation as presid... of Cuba and commander-in-chief of Cuba's milit... on Tuesday, according to a letter published by state-run newspaper Granma.

cnn 9:23:00 AM CET

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

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

iranpressnews 13:36:00 o'clock CET

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

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

kleinezeitung 10...

Fidel Castro zrezygnował! de en es fr it nl

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

Fidel Castro renuncia a la Presidencia del Consejo de Estado de... 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

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

Fidel Castro går av de en es fr it nl

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

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 že 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

هافانا، كوبا (CNN) -- ... 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

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

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.

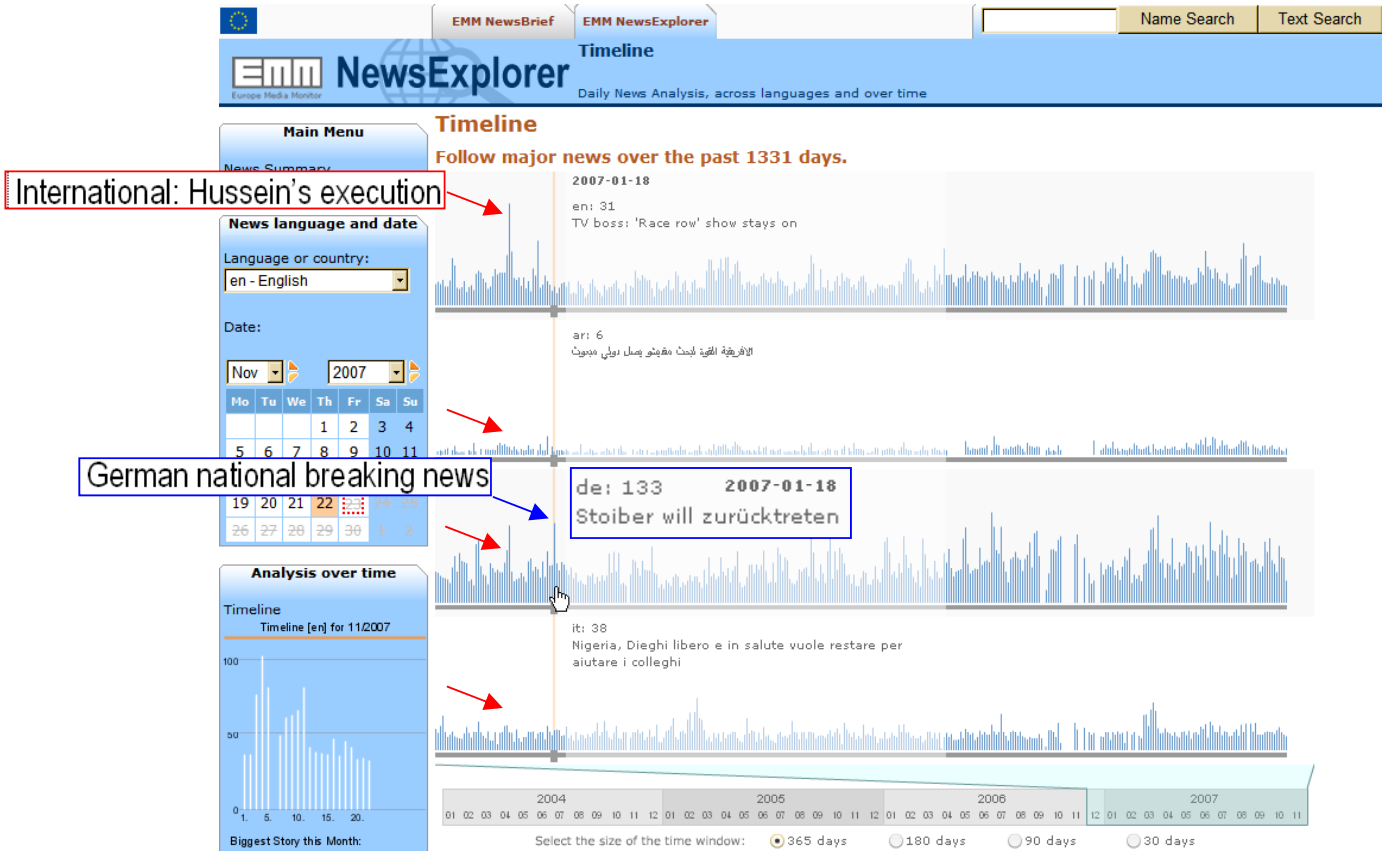
hurriyetim 10:15:00 CET

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...

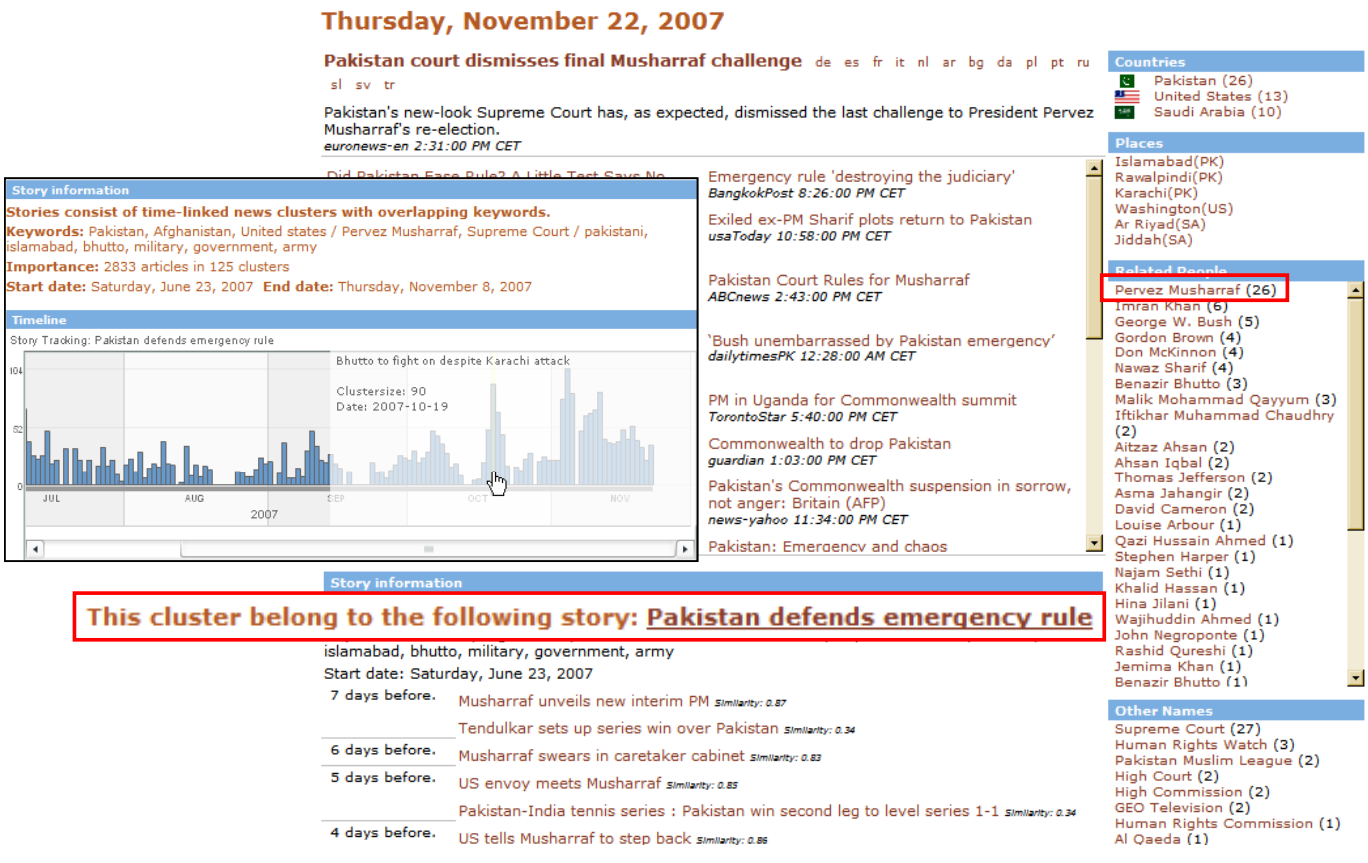
epl 23:17:00 CET

NewsExplorer – Time line: biggest clusters per day



NewsExplorer – Aggregation of clusters into longer 'stories'

live



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
(1. 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.

[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

NewsExplorer – Relation exploration

live

Associated People

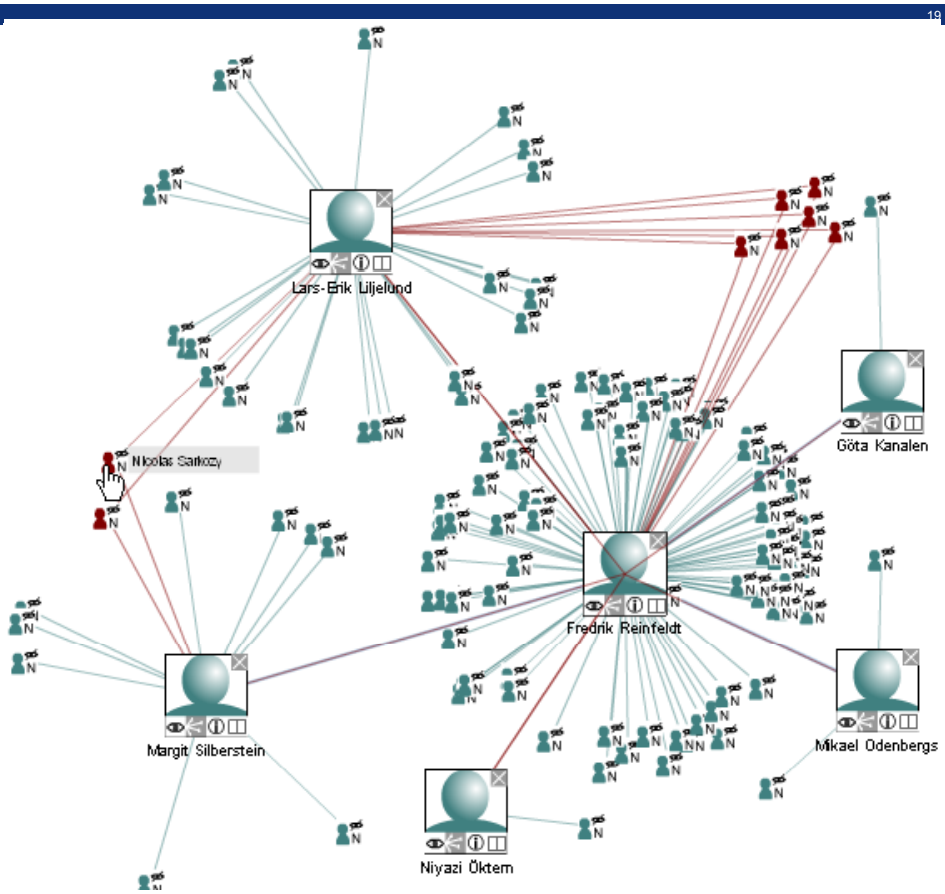
Salman Bashir (1.9)
Джон Негропонте (1.5)
Nawaz Sharif (1.2)
Раджа Омар Хатаб (1.2)
Мухоммада Али Джинны (1.2)
Зульфикара Али Бхутто (1.2)
Iftikhar Muhammad Chaudhry (1.1)
Christian College (1.1)
Tariq Azeem (1.1)
Benazir Bhutto (1.1)
Malik Mohammad Qayyum (1.0)
Chaudhry Shujaat Hussain (1.0)
Furqan 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)

Example: Pervez Musharraf & Iftikhar Chaudhry

NewsExplorer – Relation exploration

Example:

- Fredrik Reinfeldt
- Lars-Erik Liljelund
- Margit Silberstein
- …



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)
- Topic detection and tracking (TDT) (Multi-monolingual cluster similarity calculation)
- Cross-lingual cluster similarity calculation (for cross-lingual TDT).

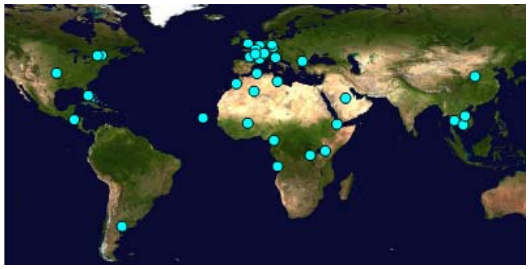
Multilinguality Motivation

Multilinguality: coverage of medical news in various languages

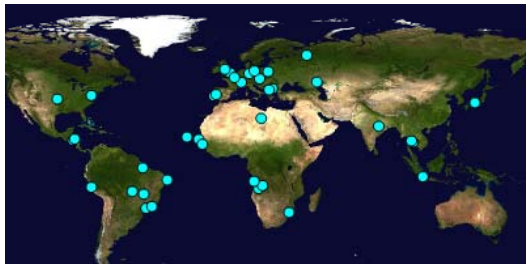
Locations mentioned in MedISys medical articles across languages – complementary coverage

Italian - German

English - French



Spanish - Portuguese



Multilinguality: More information about relations between people

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)
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

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

Key Titles and Phrases

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

External resources

Image obtained automatically from Wikipedia

Read Wikipedia entry

Steinberger Ralf & Bruno Pouliquen (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)	spaanse (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)
	presidente (es,pt - 53)

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

NER: Known person + organisation names

- 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

NER: New person + organisation names

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

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.

NER: Most frequent trigger words across languages (data: 2007)

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)

JRC EUROPEAN COMMISSION

NER: Inflection of trigger words

- 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)

Name Recognition - Evaluation results (2005)

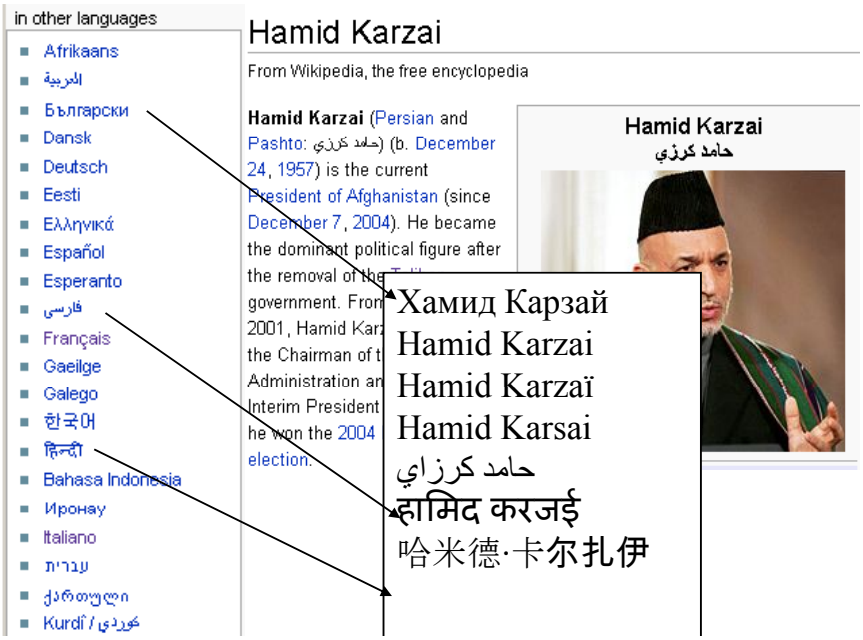
Language	*# Rules*	*# Texts*	*# Names*	*Average Precision*	*Average Recall*	*Average F-measure*
English	1100	100	405	92	84	88
French	1050	103	329	96	95	95
German	2400	100	327	90	96	93
Spanish	580	94	274	85	84	84
Italian	440	100	298	92	90	91
Russian	447	61	157	81	69	74

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.

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

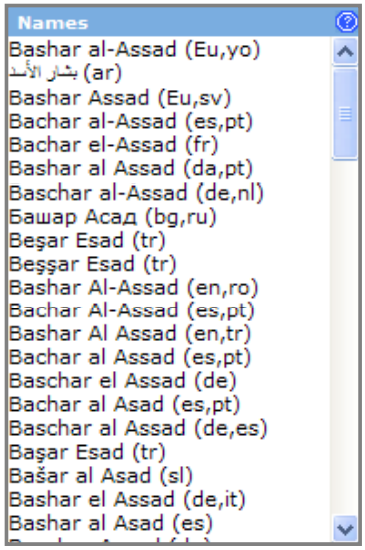
- Transliterate each character, or sequence of characters, by a Latin correspondent
  - ψ => ps
  - λ => l
  - μπ => b

  - 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
  - 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**

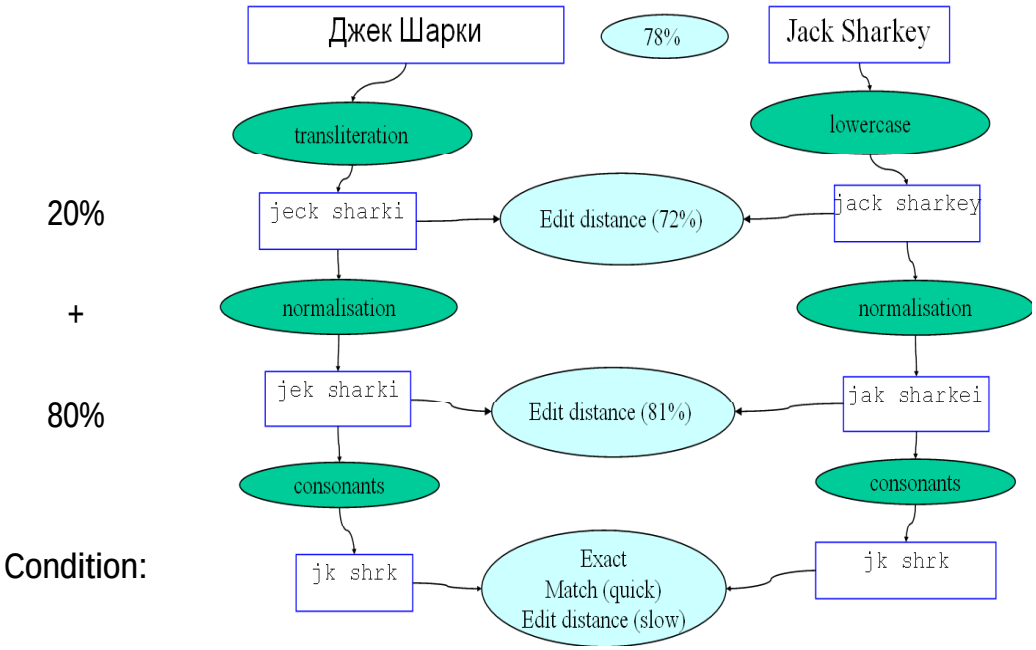
Name	Normalised form
Mohammed Siad Barre, Mohamed Siad Barré, Мохаммед Сиад Барре, محمد سياد بري	**mhmd sd br** (mohamed siad bare)
Mahmoud Ahmadinejad, Mahmūd Ahmadīnežād	**mhmd hmdnjd** (mahmud ahmadinejad)
Сергей Куприянов, Sergei Kupriyanov, Sergei Kuprianow, Sergueï Kouprianov	**srg kprnv** (sergei 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

Name variants

Ali Larijani (Eu,en)
Ali Laridschani (de)
Ali Lariyani (es)
Ali Laridžani (sl)
علي لاريجاني (ar)
Alí Larijani (es)
Ari Larijani (en)
Alí 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)
علی اردشیر لاریجانی (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 -

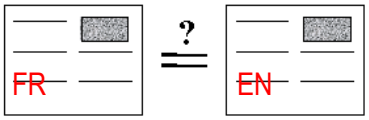
live

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

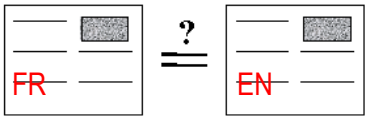
Fidel Castro announced his resignation as president of Cuba and commander-in-chief of Cuba's military on Tuesday, according to a letter published by state-run newspaper Granma.
cnn 9:23:00 AM CET

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)

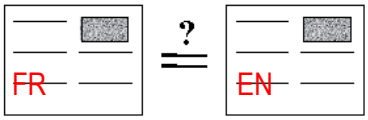
\+ Achieved results are relatively good

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

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

- EU: 22 official languages → 231 language pairs (462 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
  - 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.

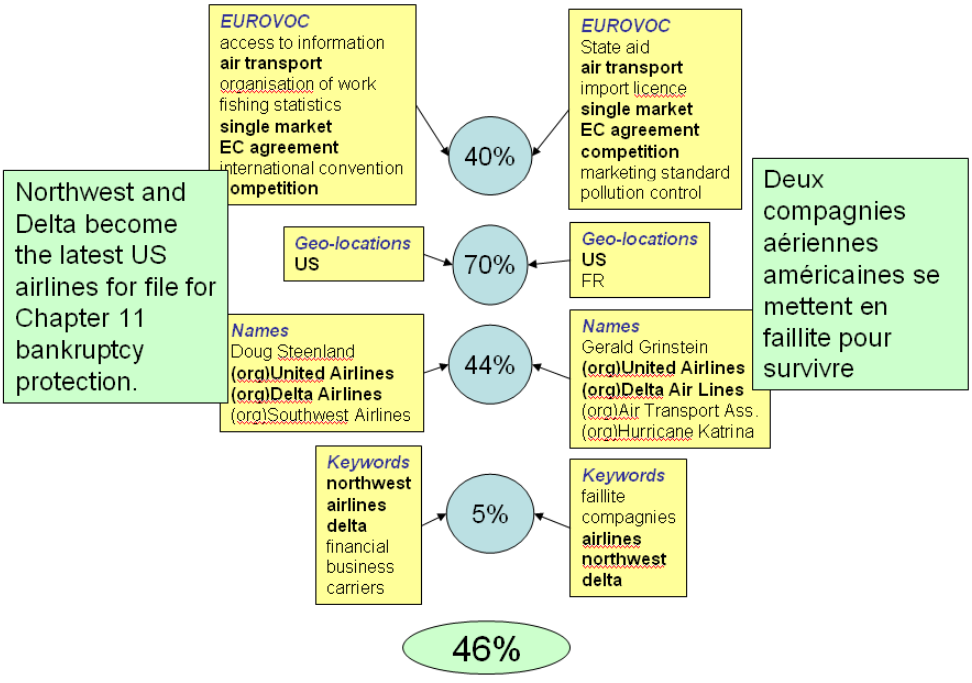
NewsExplorer - Cross-lingual cluster linking

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

- 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	*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)

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

The 'JRC-Acquis' parallel corpus in 22 languages (v. 3)

- 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

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.

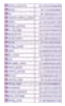
NewsExplorer - Cross-lingual cluster linking

Language-independent features for multilingual document representation

No MT or bilingual dictionaries

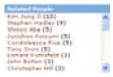
19 languages

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



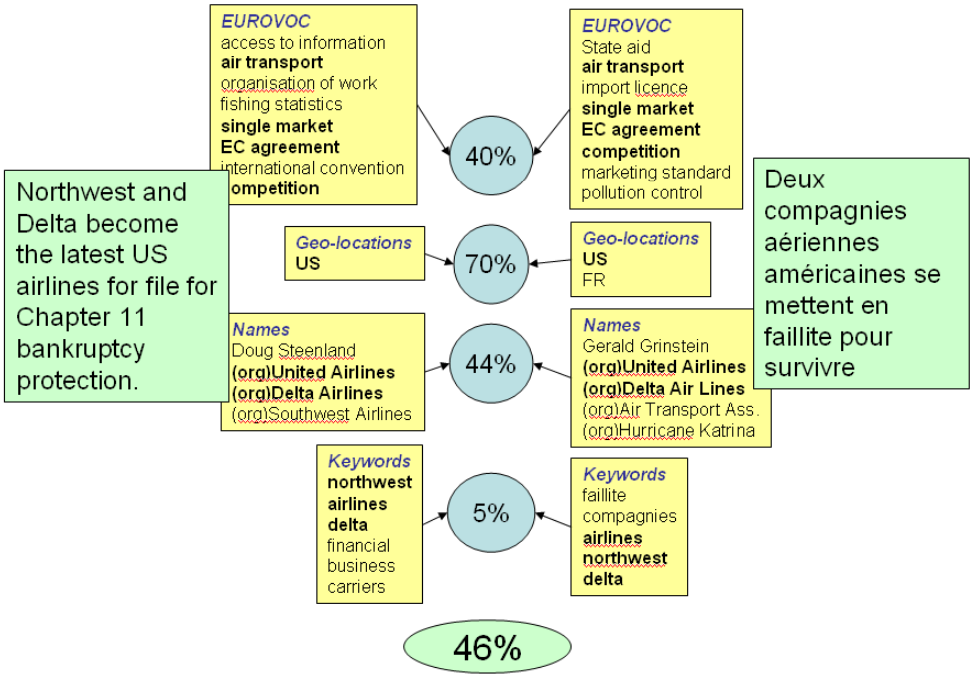
Sim1 (40%):
Multilingual Eurovoc subject domains

Sim2 (30%):
Geo-locations



Sim3 (20%):
Names + variants

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



JRC EUROPEAN COMMISSION

Summary and ongoing work

Summary

- 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.

JRC EUROPEAN COMMISSION

Ongoing work

- 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