



Cross-lingual Linking of News Clusters in Various Languages Avoiding the Usage of Bilingual Linguistic Resources

International Workshop on Intelligent Information Access
Helsinki, Finland, 8 July 2006

Ralf Steinberger, Bruno Pouliquen, Camelia Ignat, Ken Blackler, Olivier Deguernel

European Commission – Joint Research Centre (JRC)

<http://langtech.jrc.it/>

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

IIIA'2006, Slide 1

Agenda

- Major statements of this talk:
 - Present a truly multilingual application (bottleneck in text analysis applications)
 - Simple means can take you a long way
- Cross-lingual document similarity (CLDS) calculation:
 - Motivation ([NewsExplorer](http://press.jrc.it/NewsExplorer))
 - Definition
 - State-of-the-art
- Overview of our approach
- Components of the system:
 - IE: Locations
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters historically
 - Linking clusters across languages
- Future work

IIIA'2006, Slide 2

- CLIR: submit a query and retrieve documents in foreign languages
 - Typically: translation of search terms → monolingual search
 - Can be applied to an open document collection such as the WWW
- CLDS: for a given document of interest, find corresponding documents in other languages.
 - ~ Query by example
 - JRC: Document analysis → representation by various language-independent features
→ documents need processing before document similarity calculation
→ Limitation to finite document collections

- Usage of Machine Translation → monolingual document similarity (TDT-3, Leek et al. 1999)
 - Usage of bilingual dictionaries → monolingual document similarity (Wactlar 1999)
 - 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* (Fortuna et al., JRC Workshop 2005)
- + Achieved results are not bad
- Bilingual approach is restricted to a few languages:

$$\text{Language pairs} = (N^2 - N) / 2 \quad (N = \text{number of languages})$$

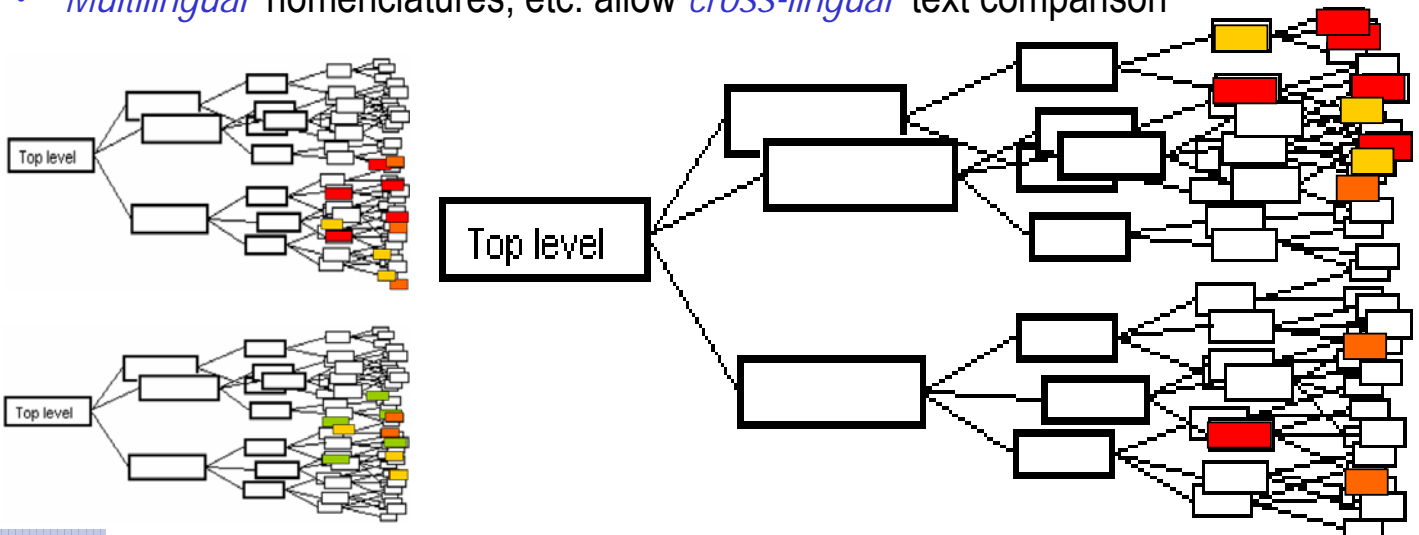
EU: 20 official languages → 190 language pairs (380 language pair directions)!

→ Attractiveness of *highly multilingual / interlingua approaches*

Steinberger, Pouliquen & Ignat:
Providing cross-lingual information access with knowledge-poor methods.
Informatica 28 (2004).

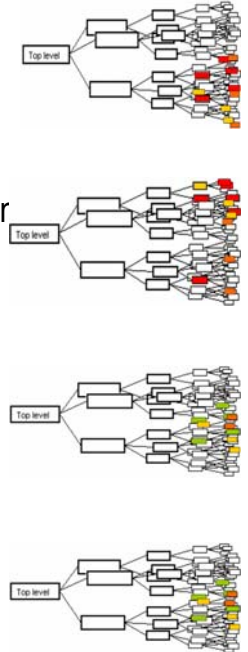
- Major statements of this talk:
 - Present a truly multilingual application (bottleneck in text analysis applications)
 - Simple means can take you a long way
- Cross-lingual document similarity (CLDS) calculation:
 - Motivation ([NewsExplorer](#))
 - Definition
 - State-of-the-art
- **Overview of our approach**
- Components of the system:
 - IE: Locations
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters historically
 - Linking clusters across languages
- Future work

- Map documents onto thesauri, nomenclatures, gazetteers, ...
- Create a text representation (*signature*) consisting of a choice of thesaurus nodes
- Relative importance of various nodes can be used
- Each mapping (geographic; medical; agricultural; ...) is one facet of a text signature
- Similar representations imply similar texts
- *Multilingual* nomenclatures, etc. allow *cross-lingual* text comparison



- Represent documents by vectors of ~language-independent features

- **Locations** (Latitude-Longitude information)
- Multilingual **thesaurus and classification system** Eurovoc
- ~ Language-independent text features
 - Normalised and merged proper **name variants** (persons and organisation)
 - Cognates and **numbers**
 - Normalised **date and currency expressions** (e.g. DATE:YYYYMMDD)
- Multilingual **nomenclatures** (products, medical terms, professions, ...)



- → CLDS (using cosine) based on this representation

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

- JRC's *Europe Media Monitor* system
 - ~ 30,000 articles per day
 - 30 languages
 - ~ 800 media news sources
 - Updated every 10 minutes
 - News in UTF8-encoded RSS format (XML)

- EMM available at <http://press.jrc.it/>
 - Breaking news
 - Topic-specific news (EU Commissioners, EU Constitution, GMO, natural disasters, communicable diseases, ...)
 - Email and SMS alerts (ca. 10,000 per day)

1. Place homographs with common words

English	
Place name	Country
And	Iran
To	Ghana
Be	India
By	Sweden

2. Place homographs with person name

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

3. Homographic place names

Place name	Number of cities with this name
Aleksandrovka	244
...	
Washington	32
London	18
Berlin	15
Paris	15

4. Completeness of gazetteer; multilinguality, e.g.

‘Санкт-Петербург’, ‘Saint Petersburg’,
‘Saint Pétersbourg’, ‘Leningrad’, ‘Petrograd’

5. Inflection

- Romanian: *Parisului* (*of Paris*)
- Estonian: *Londonit* (London),
New Yorgile (New York)
- Arabic: *نوس يرابل* (*the Paris inhabitants*)
[albaRiziu:n]

→ Usage of suffix lists to generate all variants

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

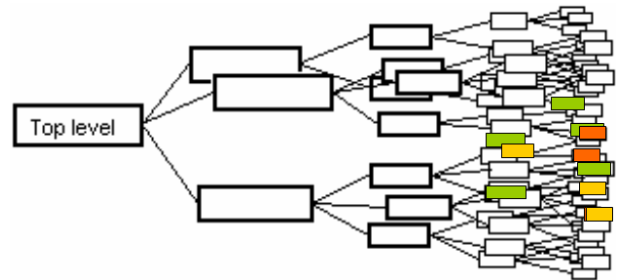
1. Geo-stop words
2. Location only if not part of person name
 - e.g. 'Kofi Annan', 'Annan'
3. Size class information
4. Country context
5. Kilometric distance
 - E.g. "from Warsaw to Brest"
 - Brest (France): 2000 km from Warsaw
 - Brest (Belarus): 200 km from Warsaw

English		French		German	
Place name	Country	Place name	Country	Place name	Country
And	Iran	De	Burkina Faso	Die	France
To	Ghana	Du	Ghana	Den	Ethiopia
Be	India	Un	Russia	Zu	Zaire

Name	City: Country
Kofi Annan	<i>Kofi</i> : Mali
	<i>Annan</i> : Scotland

Class	Explanation	Example	Weight
0	country name	Italy	80
1	capital	Rome	80
2	main city	Milan	80
3	province level	Varese	30
4	small city	Sesto Calende	20
5	village	Ispra	10
6	small settlement, hamlet	-	5

For details, see Pouliquen et al. (2006, LREC)



- List of place names found
- Frequency count per country (city, continent, region, ...)
- Frequency can be normalised, using TF.IDF or similar

- Major statements of this talk:
 - Present a truly multilingual application (bottleneck in text analysis applications)
 - Simple means can take you a long way
- Cross-lingual document similarity (CLDS) calculation:
 - Motivation ([NewsExplorer](#))
 - Definition
 - State-of-the-art
- Overview of our approach
- Components of the system:
 - IE: Locations
 - **IE: Person and Organisation names**
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters historically
 - Linking clusters across languages
- Future work

en	death of former Prime Minister Rafik Hariri, blamed by many opposition
es	asesinato del ex primer 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 pommip
ar	اغتيال رئيس الوزراء السابق رفيق الحريري بأياد يهودية وما حدث سابقا
ru	Бывший премьер-министр Ливана Рафик Харири, который

- Lookup of known names from database
 - currently 400,000 names (excluding spelling variants)
 - ~1000 new names per day
 - Pre-generate morphological variants:
 - **Tony**(a|o|u|om|em|m|ju|jem|ja)?\s+**Blair**(a|o|u|om|em|m|ju|jem|ja)
- "Guessing" names using empirically derived *lexical patterns*
 - **Trigger word(s) + Name Surname**
 - President, Minister, Head of State, Sir, American
 - "death of", "[0-9]+-year-old", ...
 - Known first names (John, Jean, Giovanni, Johan, ...)
 - Combinations: "56-year-old former prime minister **Kurmanbek Bakiyev**"

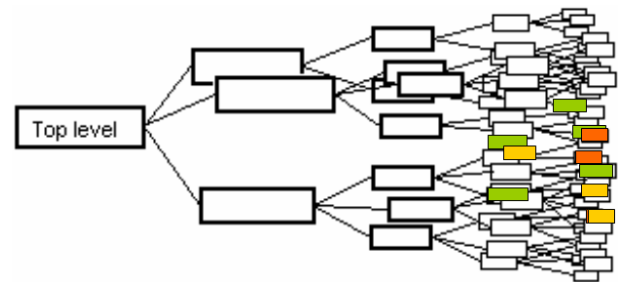
- Recognition of person names, using regular expressions (Slovene example):
 - **kandidat**(a|u|om)?
 - **legend**(a|e|i|o)
 - **milijarder**(ja|ju|jem)?
 - **predsednik**(a|u|om|em)?
 - **predsednic**(a|e|i|o)
 - **ministric**(a|e|i|o)
 - **sekretar**(ja|ju|jom|jem)?
 - **diktator**(ja|ju|jem)?
 - **playboy**(a|u|om|em)?
- + uppercase words

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

- For all new names found: apply *approximate name matching*
 - Based on sets of letter bigrams and letter trigrams
 - Merge two names if cosine similarity is > 70%
- Collect variants automatically from [Wikipedia](#)
- Cross-lingual* name matching
 - Тони Блеър => Tony Blair
 - ΙΥΙΛΑΝΤ ΑΛΛΑΟΥΙ => Iyad Allaoui
- Details: Pouliquen et al. (Journal Corela, Special Issue, 12/2005)

name	count	lang(s)
Rafik Hariri	2550	(Eu..sv)
Rafiq Hariri	827	(Eu..pl)
Rafik al-Hariri	494	(de..nl)
Rafic Hariri	315	(Eu..nl)
Rafiq al-Hariri	224	(de..en)
Rafiq Al Hariri	30	en
رديح حلا فيسر	12	ar
Rafik Al Hariri	9	en
Rafik al Hariri	9	de
Rafik el-Hariri	8	de
Rafiq Al-Hariri	6	(da..en)
Rafik Al-Hariri	6	(de..en)
Rafik Hariri Hariri	3	de
Rafik el Hariri	3	de
...

- List of numerical person IDs found
- Frequency list per name



Related People

- Kim Jong Il (10)
- Stephen Hadley (9)
- Shinzo Abe (5)
- Junichiro Koizumi (5)
- Condoleezza Rice (5)
- Tony Snow (5)
- Donald Rumsfeld (3)
- John Bolton (3)
- Christopher Hill (3)

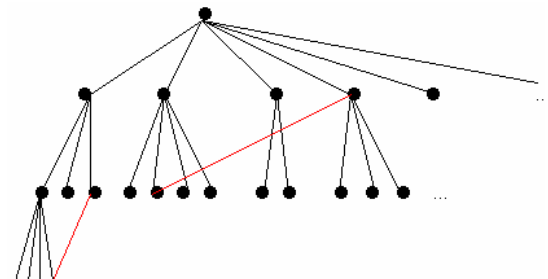
- Major statements of this talk:
 - Present a truly multilingual application (bottleneck in text analysis applications)
 - Simple means can take you a long way
- Cross-lingual document similarity (CLDS) calculation:
 - Motivation ([NewsExplorer](#))
 - Definition
 - State-of-the-art
- Overview of our approach
- Components of the system:
 - IE: Locations
 - IE: Person and Organisation names
 - **Categorisation into subject domains (Eurovoc classes)**
 - Clustering of news
 - Linking clusters historically
 - Linking clusters across languages
- Future work

- **Multilingual list of terms** (ca. 20 languages)
- Over 6000 classes
- About many different subject areas (wide coverage)
- Developed by the European Parliament and others
- Actively used **to manually index and retrieve documents** in large collections

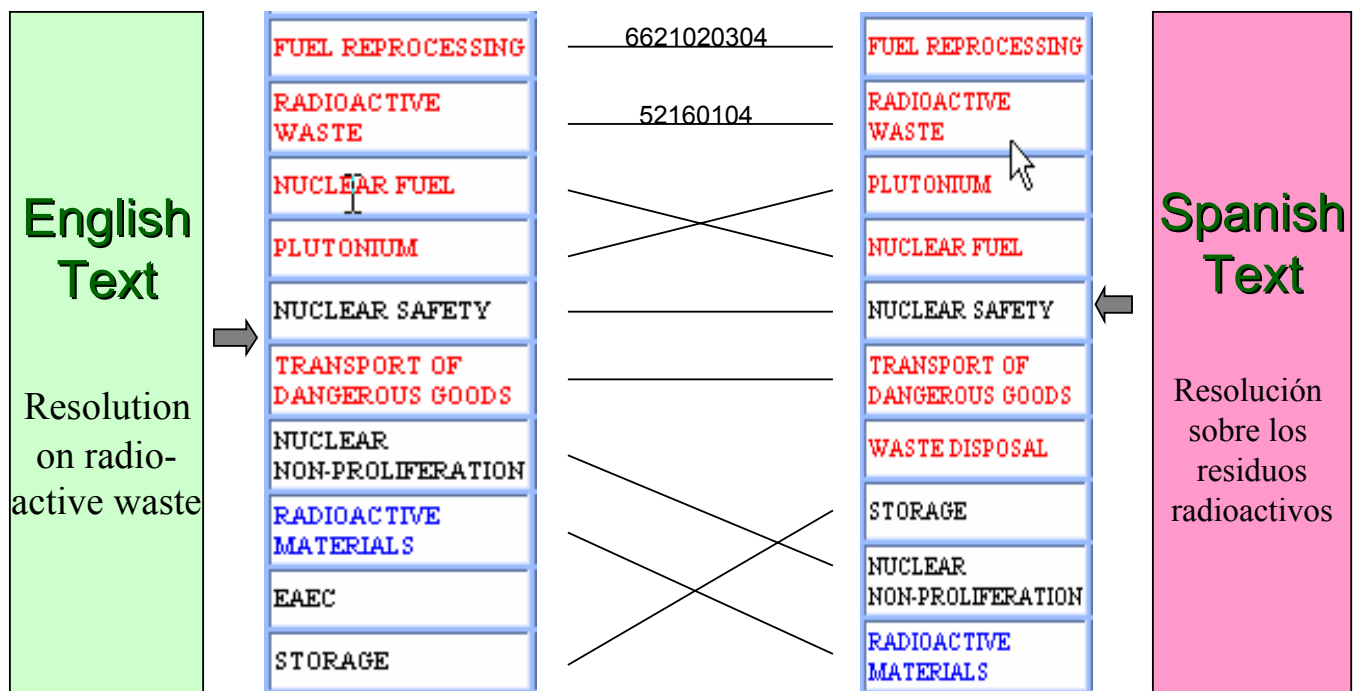
(fine-grained classification and cataloguing system)

- Hierarchically organised into up to 8 levels
Relations:

- Broader Terms
- Narrower Terms
- **Related Terms**



EN	SL
AAMS countries	države ZADM
Aarhus	Aarhus
abandoned child	zapuščeni otrok
abandoned land	opuščeno zemljišče
ABM Agreement	sporazum ABM
abolition of customs duties	odprava carin
abortion	umetna prekinitev nosečnosti
Abruzzi	Abruci
absenteeism	izostajanje z dela
absolute majority	absolutna večina
abstentionism	volilna neudeležba
Abu Dhabi	Abu Dhabi
abuse of power	zloraba pooblastil
academic freedom	akademska svoboda
access to a profession	dostop do poklica
access to Community information	dostop do informacij Skupnosti
access to education	dostop do izobraževanja
access to information	dostop do informacij



- Eurovoc is a conceptual thesaurus

E.g.

- SPORT
- PROTECTION OF MINORITIES
- CONSTRUCTION AND TOWN PLANNING
- RADIOACTIVE MATERIALS

- *Keyword assignment vs. term extraction*

- *6000 classes*
- *Unevenly distributed*
- *Heterogeneous training text types*
- *Multi-label categorisation*

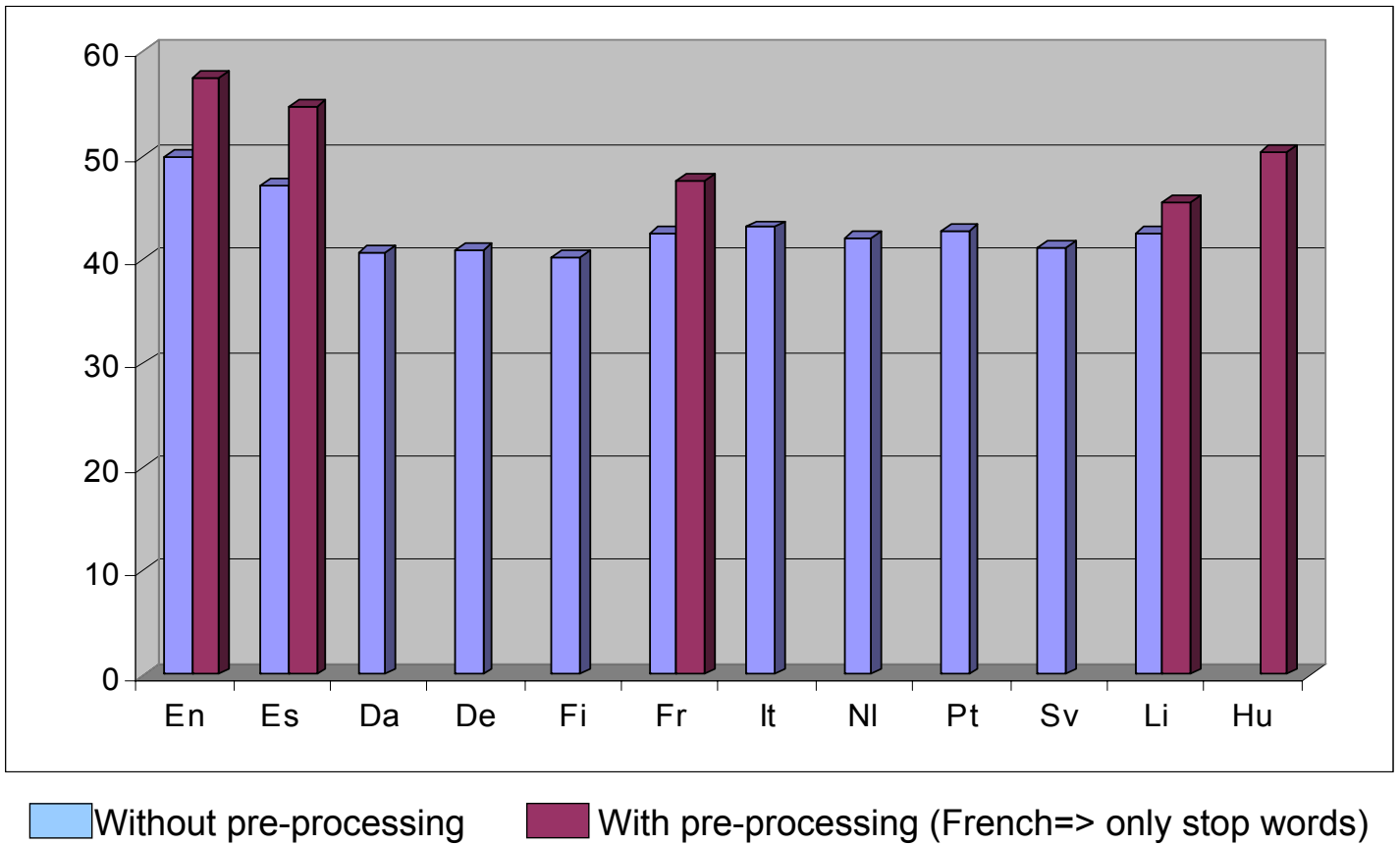
- Profile-based, category ranking task
 - Training: Identification of most significant words for each class
 - Assignment: combination of measures to calculate similarity between profiles and new document
- Empirical refinement of parameter settings
 - Training:
 - Stop words
 - Lemmatisation
 - Multi-word terms
 - Consider number of classes of each training document
 - Thresholds for training document length and number
 - Methods to determine significant words per document (log-likelihood vs. chi-square, etc.)
 - Choice of reference corpus
 - ...
 - Assignment:
 - Selection and combination of similarity measures (cosine, okapi, ...)
 - ...
- See Pouliquen et al. (Eurolan 2003)

Sample profile: RADIOACTIVE MATERIALS

deuterium	35.7836791092845
lithium	33.0805724769899
thorium	32.560703225522
tritium	32.0826451843048
nuclear_material	13.79399100837
radioactive_material	7.84970673161556
plutonium	6.72955494180221
radioactive_substance	6.43422856440347
nuclear	5.851612117697
undine_uta_bloch_von_blottnitz	5.53278869694883
radioactive	4.89399300382035
nuala_ahern	4.04706620369489
radon	4.03336435560442
mox	3.5654196472221
uranium	3.33954480260962
illegal_traffic	3.03072833135354

Sample profile: FISHERY MANAGEMENT

fishery-related	fishery_resource	54.4721542368385
	fishing	49.111563204862
	fish	46.196436023147
	common_fishery_policy	44.6741845971235
	fishery	44.1911518447189
	fishing_activity	43.3777671334009
	fly_the_flag	42.8744724542378
	aquaculture	39.2749719215554
	conservation	38.3480454820621
	vessel	37.911138722495
management-related	fishing_vessel	37.8343365844963
	catch	36.8503034704154
	fish_stock	34.5283935973103
	tacs	34.388453583343
	allowable_catch	33.2880590561664
	catch_quota	32.2683540654092
	control_system	31.1753892078216
	fish_for	29.8386698340017
	nautical_mile	29.541061528168
	fishing_right	29.1916760888221
centimetre	28.7167313169535	
control_measure	28.0527345432075	
gross_tonnage	28.0043616725124	
fishing_zone	27.8678836557192	



Correct descriptors

compared to performance of professional human indexers

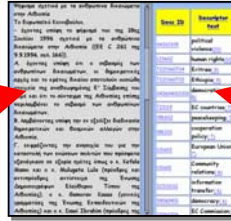
English: **83%**

Spanish: **80%**

Desc ID	Descriptor text	B	S	?	bt	nt	G	V	Comment
72060411	Reino Unido(26)								victoria
56260205	bovino(25)								victoria
56310601	enfermedad animal(22)								victoria
60360104	sacrificio de animales(21)								victoria
284104041103	riesgo sanitario(17)								victoria
5611010501	prima por sacrificio voluntario de reses(17)								I think this unsuitable. ... usual way to 08-APR-03 victo
20260102	protección del consumidor(17)								victoria
56060104	legislación veterinaria(17)								victoria
5606010401	inspección veterinaria(15)								victoria
1006020102	presidencia del Consejo CE(15)								victoria
2841040403	medicina preventiva(14)								victoria
6011010604	carne bovina(14)								victoria

Analysis

- Danish
- Dutch
- English
- Finnish
- French
- German
- (Greek)
- Italian
- Portuguese
- Spanish
- Swedish
- (Hungarian)



Display

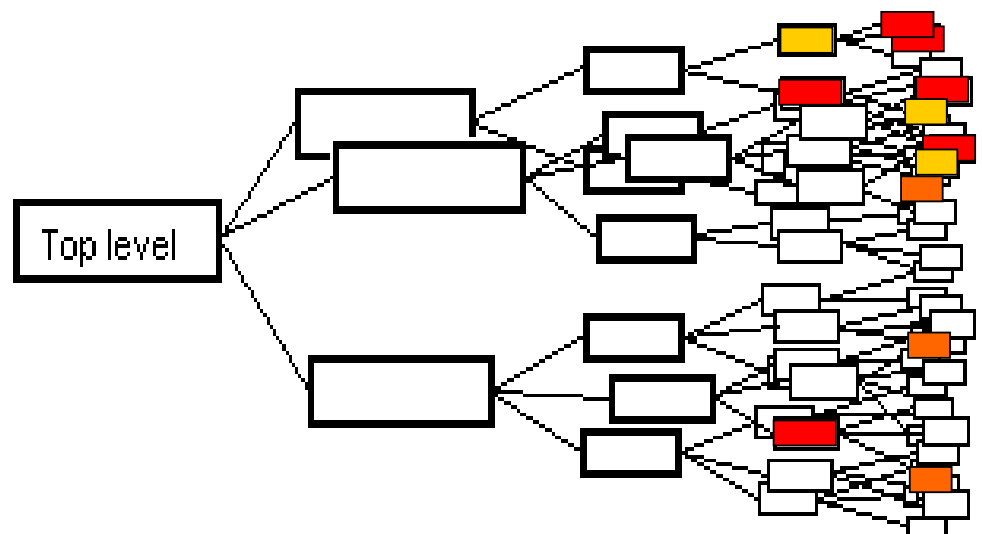
- Danish
- Dutch
- English
- Finnish
- French
- German
- Greek
- Italian
- Portuguese
- Spanish
- Swedish
- Czech
- Croatian
- Hungarian
- Latvian
- Lithuanian
- Polish
- Romanian
- Slovak
- Slovene

Available also in

- Albanian
- Russian

- Ranked list of Eurovoc descriptor codes found for each document

Descriptor ID	Cosine
5641040706000000	0.360
5641020000000000	0.308
5641040200000000	0.280
5641040100000000	0.279
5641040700000000	0.270
5641040704000000	0.261
5641040101000000	0.253
5641040600000000	0.252
5206040100000000	0.251
5641050000000000	0.232
5641040800000000	0.213
5641040000000000	0.203
5641040705000000	0.181
5641060100000000	0.179
5641010000000000	0.176
5641040201000000	0.176
...	



- Available at: <http://langtech.jrc.it/JRC-Acquis.html>
 - Steinberger et al. (2006, LREC)
 - Average of 8.8 Million words per language
 - Pair-wise alignment for all 210 language pairs
 - Average of 7600 documents per language
 - Most documents have been Eurovoc-classified manually
- useful for:
1. Training of multilingual subject domain classifiers.
 2. Production of multilingual lexical space (LSA, KCCA)
 3. Training of automatic systems for Statistical Machine Translation.
 4. Producing multilingual lexical or semantic resources such as dictionaries or ontologies.
 5. Training and testing multilingual information extraction software.
 6. Automatic translation consistency checking.
 7. Testing and benchmarking alignment software (sentences, words, etc.), across a larger variety of language pairs.
 8. All types of multilingual and cross-lingual research.

- Major statements of this talk:
 - Present a truly multilingual application (bottleneck in text analysis applications)
 - Simple means can take you a long way
- Cross-lingual document similarity (CLDS) calculation:
 - Motivation ([NewsExplorer](#))
 - Definition
 - State-of-the-art
- Overview of our approach
- Components of the system:
 - IE: Locations
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - **Clustering of news**
 - Linking clusters historically
 - Linking clusters across languages
- Future work

- Vector of keywords and their keyness using log-likelihood test (Dunning 1993)
 - alternatives: TF.IDF, chi-square, ...
 - comparing word frequency in text with word frequency in comparable reference corpus

“Michael Jackson Jury Reaches Verdicts”

Keyness	Keyword	Keyness	Keyword
109.24	jackson	9.39	verdict
41.54	neverland	7.56	testimony
37.93	santa	6.50	maria
32.61	molestation	4.09	michael
24.51	boy	1.73	reached
24.43	pop	1.68	ap
20.68	documentary	1.05	appeared
18.79	accuser	0.53	child
13.59	courthouse	0.50	trial
11.12	jury	0.45	monday
10.08	ranch	0.26	children
9.60	california	0.09	family

Monday, June 13, 2005

Michael Jackson Jury Reaches Verdicts es de it nl

Jackson, 46, was accused of molesting the then-13-year-old boy and plying him with wine at the pop star's Neverland ranch in 2003. Jackson had befriended the boy, a cancer survivor, and they appeared together when Jackson was interviewed for the documentary "Living With Michael Jackson." *ABCnews 13/06/2005 21:54*

Jackson cleared of all 18 charges
ananova 13/06/2005 23:36

Timetable Of Events Which Led To Court
skynews 13/06/2005 23:30

Week 2 Of Jackson Deliberations
CBSnews 13/06/2005 12:07

Jackson jurors set for second week
NEWScomAU 13/06/2005 03:52

Jackson jury into second week
TheAustralian 13/06/2005 18:12

Music's misunderstood superstar
Bbc 13/06/2005 23:25

The Extraordinary Life Of A Music Icon
five 13/06/2005 23:32

Jury finds Michael Jackson innocent of all charges
irishtimes 13/06/2005 23:50

Michael Jackson found not guilty on all charges
itv 13/06/2005 23:34

JACKSON NOT GUILTY
skynews 13/06/2005 23:22

Michael Jackson cleared of abuse
abc 13/06/2005 22:25

Analysis: Testing times ahead
guardian 13/06/2005 23:48

Jackson arrives at court
TheAustralian 13/06/2005 21:27

- Aim: show to what extent a text talks about a certain country?
 - some countries are generally talked about much more than others
 - normalise the frequency count by average occurrence frequency
 - using the log-likelihood test
- Add country score vector to keyword vector

10.4184	*us*
1.5610	*gb*
1.5610	*il*
1.5610	*br*

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

Monday, June 13, 2005

Michael Jackson Jury Reaches Verdicts es de it nl

Jackson, 46, was accused of molesting the then-13-year-old boy and plying him with wine at the pop star's Neverland ranch in 2003. Jackson had befriended the boy, a cancer survivor, and they appeared together when Jackson was interviewed for the documentary "Living With Michael Jackson." *ABCnews 13/06/2005 21:54*

Jackson cleared of all 18 charges
ananova 13/06/2005 23:36

Timetable Of Events Which Led To Court
skynews 13/06/2005 23:30

Week 2 Of Jackson Deliberations
CBSnews 13/06/2005 12:07

Jackson jurors set for second week
NEWScomAU 13/06/2005 03:52

Jackson jury into second week
TheAustralian 13/06/2005 18:12

Music's misunderstood superstar
Bbc 13/06/2005 23:25

The Extraordinary Life Of A Music Icon
five 13/06/2005 23:32

Jury finds Michael Jackson innocent of all charges
irishtimes 13/06/2005 23:50

Michael Jackson found not guilty on all charges
itv 13/06/2005 23:34

JACKSON NOT GUILTY
skynews 13/06/2005 23:22

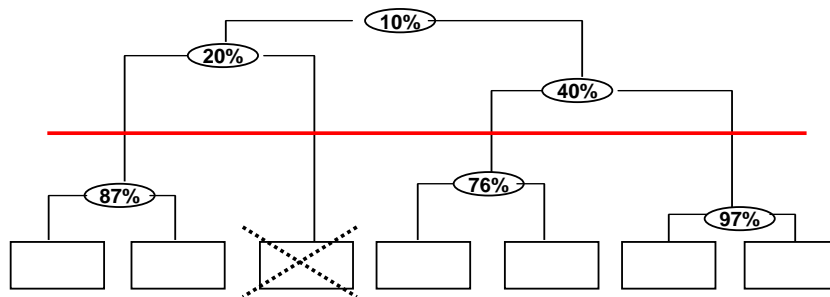
Michael Jackson cleared of abuse
abc 13/06/2005 22:25

Analysis: Testing times ahead
guardian 13/06/2005 23:48

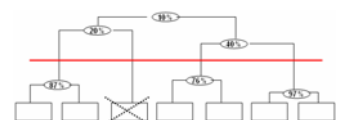
Jackson arrives at court
TheAustralian 13/06/2005 21:27

- Input: Vectors consisting of keywords and country score
- Similarity measure: cosine
- Method: Bottom-up group average unsupervised clustering
- Build the hierarchical clustering tree (dendrogram)
 - Retain only “big” nodes in the tree with a high cohesion (minimum intra-node similarity 45%)

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



- Details: Pouliquen et al. (CoLing 2004)
- Evaluation results depending on similarity threshold



Similarity threshold	“Precision”	“Recall”
15%	88%	100%
20%	92%	98%
40%	98%	86%
60%	99%	78%
80%	99%	67%

- For each cluster, find the most representative article (*medoid*)
- Use its title as the title for the cluster

Monday, June 13, 2005

Michael Jackson Jury Reaches Verdicts es de it nl

Jackson, 46, was accused of molesting the then-13-year-old boy and plying him with wine at the pop star's Neverland ranch in 2003. Jackson had befriended the boy, a cancer survivor, and they appeared together when Jackson was interviewed for the documentary "Living With Michael Jackson."
ABCnews 13/06/2005 21:54

Jackson cleared of all 10 charges
ananova 13/06/2005 23:36

Timetable Of Events Which Led To Court
skynews 13/06/2005 23:30

Week 2 Of Jackson Deliberations
CBSnews 13/06/2005 12:07

Jackson jurors set for second week
NEWScomAU 13/06/2005 03:52

Jackson jury into second week
TheAustralian 13/06/2005 18:12

Music's misunderstood superstar
bbc 13/06/2005 23:25

The Extraordinary Life Of A Music Icon
five 13/06/2005 23:32

Jury finds Michael Jackson innocent of all charges
irishtimes 13/06/2005 23:50

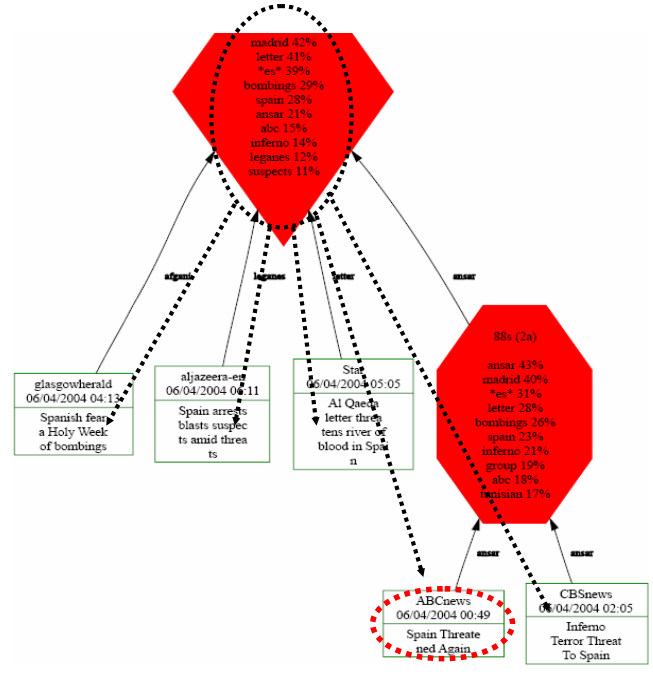
Michael Jackson found not guilty on all charges
rtv 13/06/2005 23:34

JACKSON NOT GUILTY
skynews 13/06/2005 23:22

Michael Jackson cleared of abuse
bbc 13/06/2005 23:25

Analysis: Testing times ahead
guardian 13/06/2005 23:48

Jackson arrives at court
TheAustralian 13/06/2005 23:27



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

- Ranked list of Eurovoc classes (40%)
- Country score (30%)
- Names + frequency (20%)
- Monolingual cluster representation *without country score* (10%)

Descriptor ID	Cosine
I5641040706000000	0.360
I5641020000000000	0.308
I5641040200000000	0.280
I5641040100000000	0.279
I5641040700000000	0.270
I5641040704000000	0.261
I5641040101000000	0.253
I5641040600000000	0.252
I5206040100000000	0.251
I5641050000000000	0.232
I5641040800000000	0.213
I5641040000000000	0.203
I5641040705000000	0.181
I5641060100000000	0.179
I5641010000000000	0.176
I5641040201000000	0.176

+ **10.4184** *us*
1.5610 *gb*
1.5610 *il*
1.5610 *br*

Related People
Kim Jong Il (10)
Stephen Hadley (9)
Shinzo Abe (5)
Junichiro Koizumi (5)
Condoleezza Rice (5)
Tony Snow (5)
Donald Rumsfeld (3)
John Bolton (3)
Christopher Hill (3)

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
10.0838	ranch
9.6021	california

- Evaluation results depending on similarity threshold
- Ingredients: 40/30/30 (names not yet considered)
- Details: Pouliquen et al. (CoLing 2004)
- Evaluation for EN → FR and EN → IT (136 EN clusters)

Similarity threshold	EN → FR Precision	EN → FR Recall *	EN → IT Precision	EN → IT Recall *
30%	84%	99%	71%	97%
60%	98%	46%	98%	42%

* Recall at 15% similarity threshold = 100%

- Major statements of this talk:
 - Present a truly multilingual application (bottleneck in text analysis applications)
 - Simple means can take you a long way
- Cross-lingual document similarity (CLDS) calculation:
 - Motivation ([NewsExplorer](#))
 - Definition
 - State-of-the-art
- Overview of our approach
- Components of the system:
 - IE: Locations
 - IE: Person and Organisation names
 - Categorisation into Subject domains (Eurovoc classes)
 - Clustering of news
 - Linking clusters historically
 - Linking clusters across languages
- **Future work**

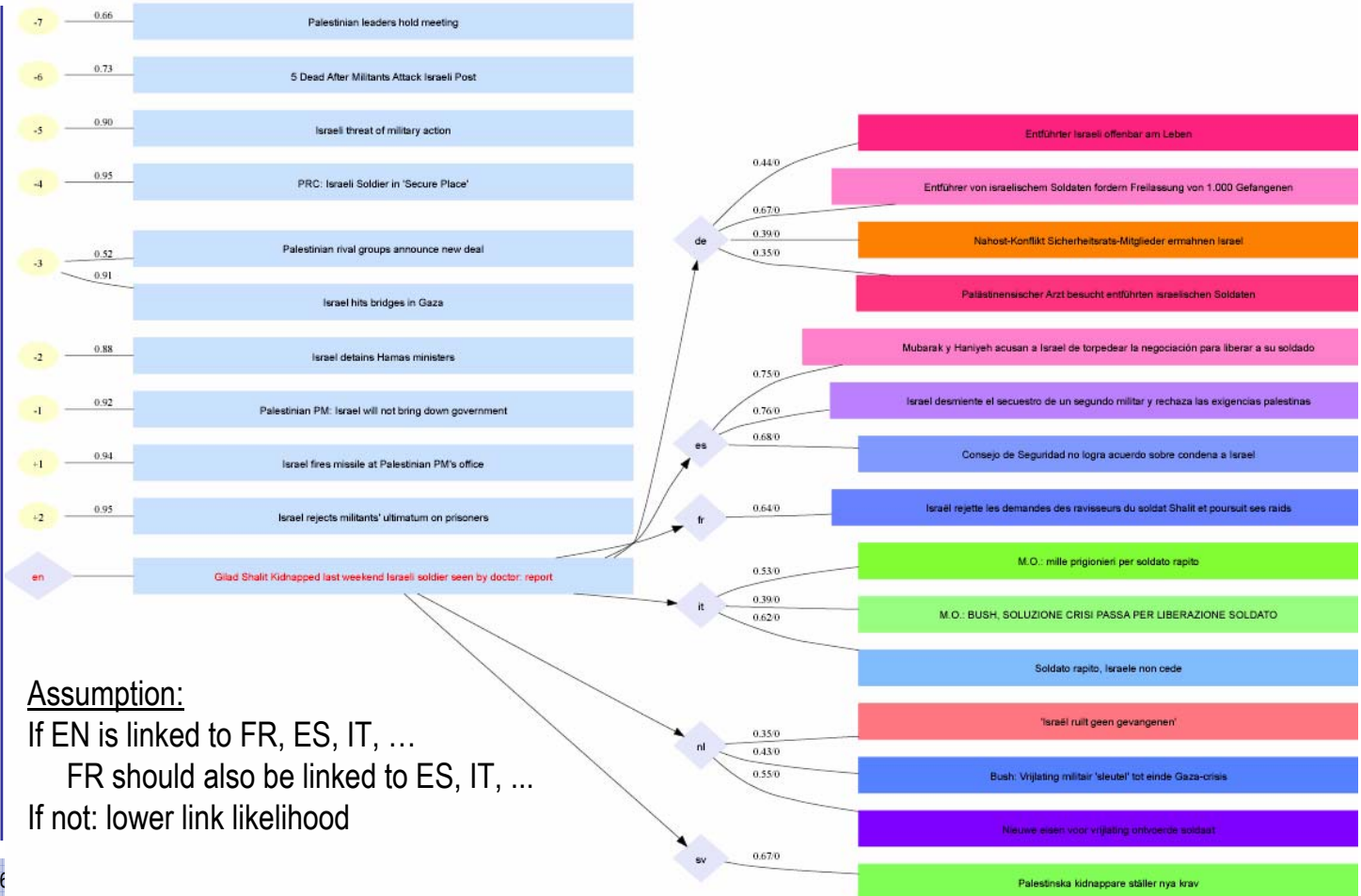
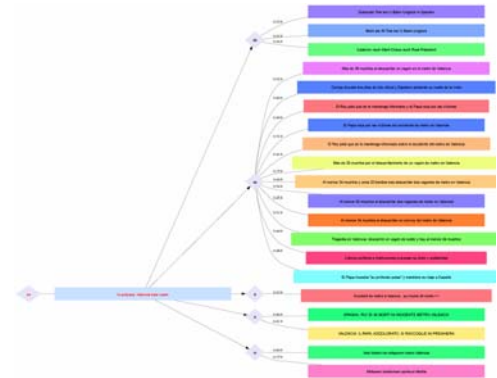
1. Add more information facets

- Dates
- Professions
- Expressions of measurement
- Vehicles
- ...

2. Empirically optimise weighting of ingredients

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

- Short-term solution: filtering bad links using heuristics based on multilingual cross-linking



Assumption:

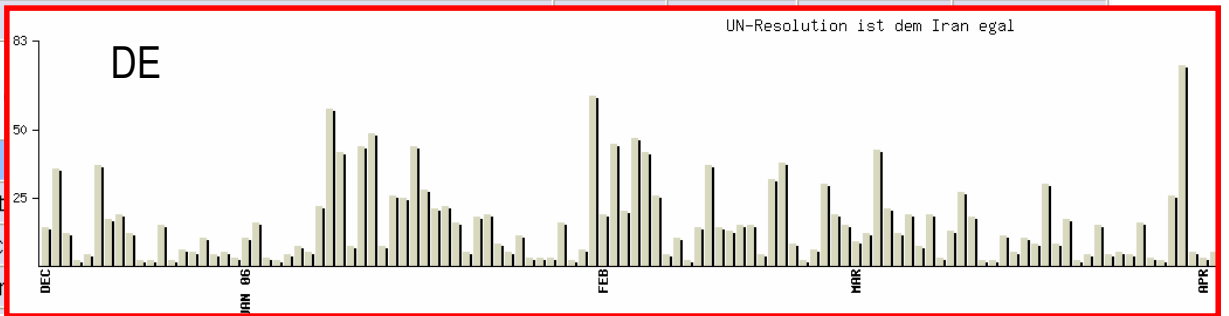
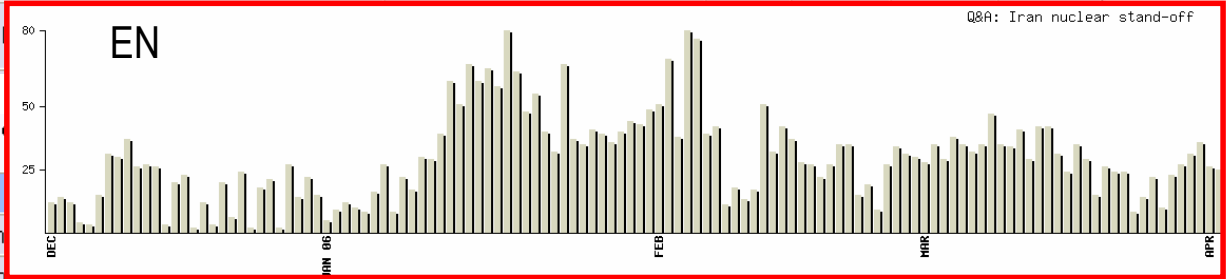
If EN is linked to FR, ES, IT, ...

FR should also be linked to ES, IT, ...

If not: lower link likelihood

Biggest stories since December 2005 [5](#) [10](#) [20](#) [30](#) [50](#)

Title	Biggest Title	Weight	Clusters	Starts	Ends
<input type="checkbox"/> Iran Prez: Move Israel To Europe info	Q&A: Iran nuclear stand-off	6506	220	08-DEC-05	05-JUL-06
<input type="checkbox"/> Bush Attempts Hard Sell on Iraq Progress info	Most-Wanted Iraq Terrorist Al-Zarqawi Dead	6343	329	01-DEC-05	05-JUL-06
<input type="checkbox"/> Four killed in					



New stories

- Title
- In pictures: Valen
- New al Qaeda no
- N.J. government

New stories

- Title
- Macedonians vot
- Stable west Afric
- Top Italian spy ar

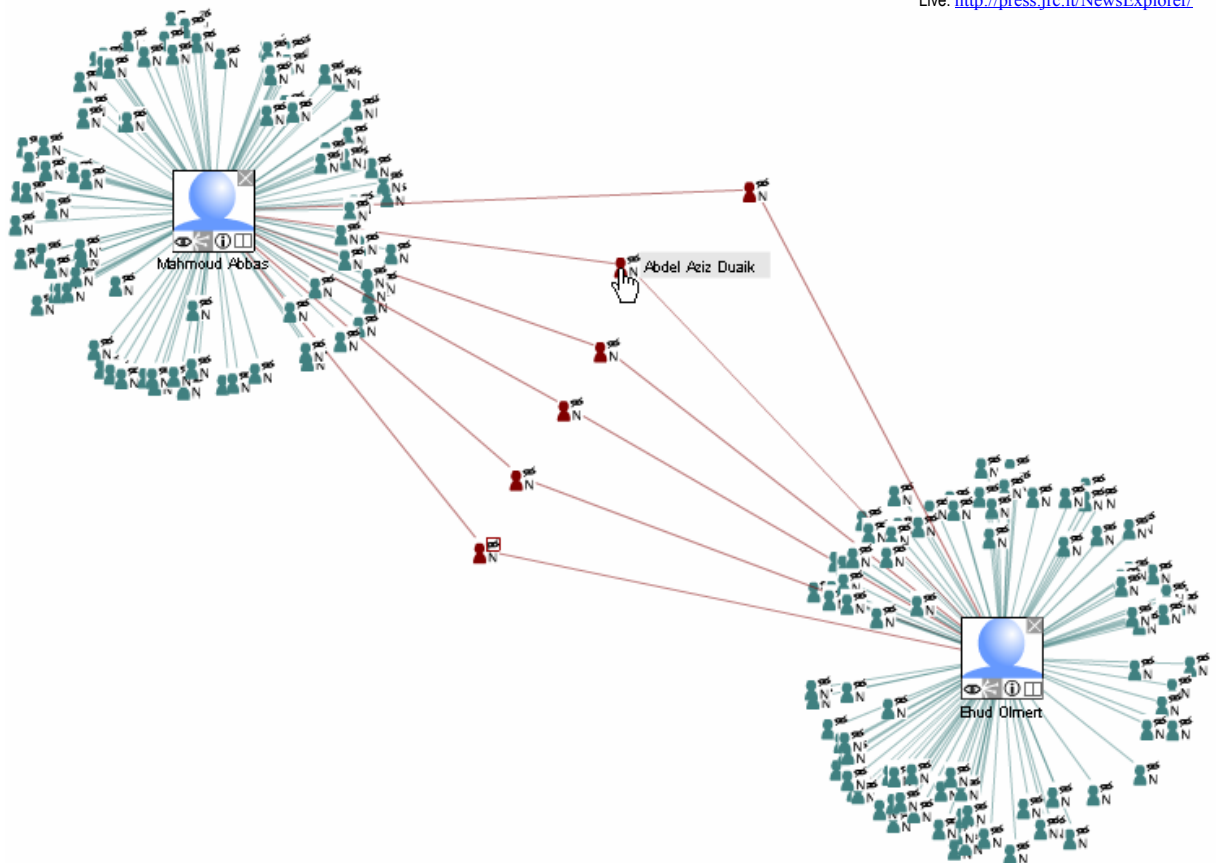
Timeline

Follow major news over the past 827 days.



Planned work (5): Linking extracted person-person relationships with multilingual stories

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



IIIA'2006, Slide 49

Conclusion

- Cross-lingual linking of documents/clusters via language-independent representations is feasible
- Simple means can take you a long way
 - JRC effort to add a new language is 1 - 6 months
 - Simple lexical patterns
 - Heuristics
 - Statistics
 - Machine Learning

IIIA'2006, Slide 50