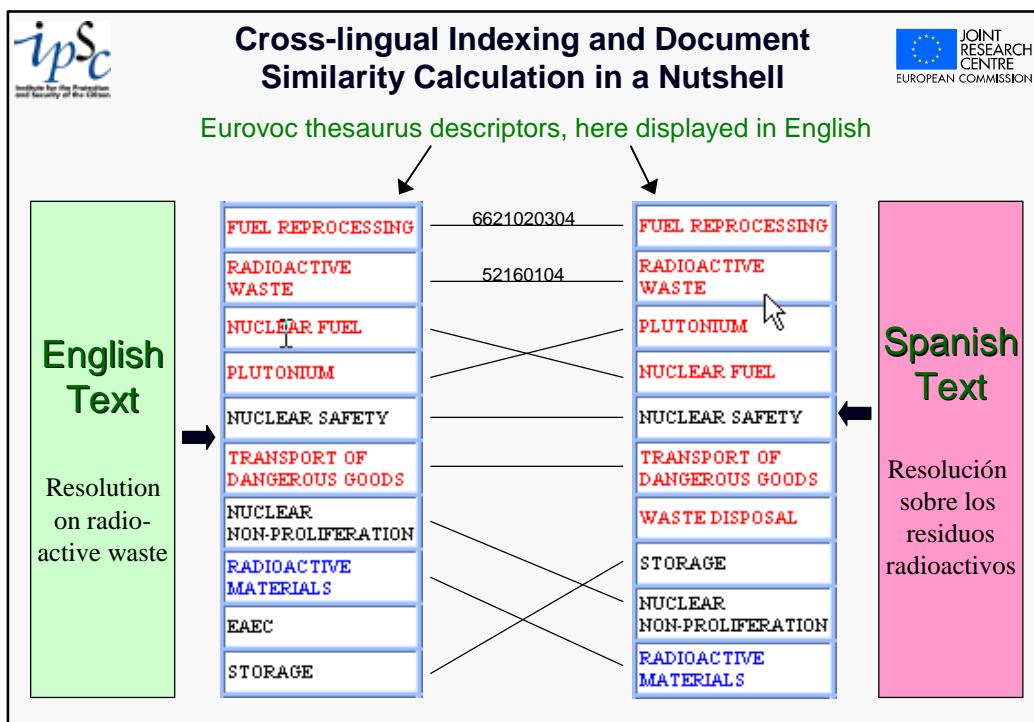



## A tool set to retrieve and analyse multilingual texts and to give users cross-lingual information access

**Université Libre de Bruxelles**  
**Séminaire: Questions Actuelles d'informatiques**  
**11 March 2003**


**Ralf Steinberger, Bruno Pouliquen, António Ribeiro, Camelia Ignat**  
 European Commission – **J**oint **R**esearch **C**entre (**JRC**)  
 Institute for the **P**rotection and **S**ecurity of the **C**itizen (**IPSC**)

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







## Agenda



- **Introduction: Who we are and what we do**
  - Joint Research Centre
  - Language Technology in IPSC
    - Document retrieval
    - Text analysis and information extraction
    - Visualisation of textual information
- **Assignment of descriptors of the multilingual thesaurus Eurovoc to texts**
  - Method; Challenges and difficulties
  - Applications:
    - Cross-lingual indexing
    - Multilingual document maps
    - Cross-lingual document similarity calculation
    - Subject-specific summarisation
- Possibilities for collaboration with the JRC



## JRC Sites in Europe



Map of Europe showing JRC sites: itm (Ireland), ie (Ireland), itu (Italy), ipsc (Belgium), ihp (Cyprus), ips (Spain), ies (Spain).

Photograph of the Ispra Site




## Joint Research Centre (JRC / CCR / GFS)






- Directorate General (DG) of the European Commission
- > 1500 scientists and technicians, ca. 2500 people
- Ispra: ca. 1800 people
- Scientific research and scientific services for DGs
- Wide range of subjects:
  - nuclear safety
  - environment (alternatives to animal testing, recognition of adulterated wine, (non-)biological food, ...)
  - Security of food and chemical products
  - dependability of information systems
  - ...
- **Multi-disciplinary** work. We mainly collaborate with *Web Technology*, *Machine Learning* and *Statistics* sectors.



## The Language Technology Team




**Current team**

Ralf Steinberger  
Bruno Pouliquen  
António Ribeiro  
Camelia Ignat


**Past members**

(Johan Hagman)  
(Stefan Scheer)  
(Marco Palazzini)  
(Giovanni Valerio)

**Small team; many languages to deal with  
=> usage of mainly statistical methods**



## Goal of JRC's Language Technology work



- **Retrieval of potentially relevant texts** (e.g. from the internet) in a variety of languages, using agent technology: [OSILIA project (2000), IDoRA for OLAF (2002/03), Breaking News – Detection and Visualisation (2003)]
- **Text analysis** and extraction of a variety of information aspects from texts; when possible: language-independent representation of the contents
  - key words (monolingual free indexing terms and **cross-lingual Eurovoc descriptors**)
  - language of texts
  - references to geographical places (and to dates)
  - (references to people, to products, etc.)
  - summary
  - Calculation of the similarity of texts; find related documents, even **across languages**
  - clustering and classification of documents
- **Visualisation of the contents**
  - of individual documents in *document profiles*
  - of whole text collections in *document maps*
  - of extracted geographical information in maps



## Retrieval of Online Newspaper Articles

OSILIA: <http://www.jrc.it/langtech/osilia.html> (year 2000)



- Automatic, daily search of ca. 350 online newspapers and other news sites (about 200 for En, Fr and De alone) on various subject areas, e.g. 'internet abuse'
- Collaboration with JRC's *Web Technology Sector* (<http://wt.jrc.it>)

- Including automatic
  - cleaning of the retrieved web pages
  - filtering
  - duplicate identification
  - relevance ranking
  - classification
  - keyword assignment
- Interface for browsing, searching, visualising the resulting collection
- Evaluation against manual newspaper clipping service was successful.



**Web-Site independent 'Cleaning' of Web Pages**

The screenshot shows a news article from 'L'Espresso' dated November 12, 2002, titled 'La lunga giornata del presidente della Camera e lo scudone a pezzi nel partito reggiano di Jada Berlusconi'. The article discusses the political situation in Italy, mentioning the resignation of Massimo D'Alema and the formation of a new government. The website interface includes a search bar, navigation links, and a sidebar with various news snippets.

**(Near-) Duplicate Identification**

The screenshot displays two news articles from 'The Wall Street Journal' dated November 12, 2002, both titled 'Philip Morris stubs out earnings forecast'. The articles report that Philip Morris has abandoned its 2003 earnings forecast, citing a sharp decline in cigarette sales and the impact of the 1998 Master Settlement Agreement. A table below compares the two articles, highlighting their similarities and differences.

"Philip Morris stubs out earnings forecast"	"Philip Morris abandons 2003 forecast"
Philip Morris abandoned its 2003 earnings forecast Tuesday in New York. Published November 12, 2002 22:00. Last Updated November 12, 2002 22:16. On Tuesday, Philip Morris abandoned its 2003 earnings forecast after warning that third-quarter earnings would be lower than expected.	Philip Morris abandoned its 2003 earnings forecast Tuesday in New York. Published November 12, 2002 22:08. Last Updated November 12, 2002 22:17. The tobacco giant said it was abandoning its 2003 earnings forecast after warning that third-quarter earnings would be lower than expected.
Shares of the world's largest tobacco company tumbled almost 14 percent to \$37.33 after it said it was not in the position to confirm its 2003 earnings projections of 10 to 15 percent growth in underlying earnings per share next year.	The world's largest tobacco company said it was not in the position to confirm its 2003 earnings projections of 10 to 15 percent growth in underlying earnings per share next year.
The Wall Street Journal's November 12, 2002, article is identical to the one.	The Wall Street Journal's November 12, 2002, article is identical to the one.
The warning to a Morgan Stanley investors conference came six weeks after Philip Morris said earnings growth this year would be much lower than its original expectations of at least 10 percent.	The warning to a Morgan Stanley investors conference came six weeks after Philip Morris said earnings growth this year would be much lower than its original expectations of at least 10 percent.
It comments also hit shares, which were down 11.4 percent at \$37.03 and UST, the biggest US maker of snuff and chewing tobacco, down 7.5 percent at \$29.93.	Philip Morris shares closed at \$37.33 and UST, the biggest US maker of snuff and chewing tobacco, down 7.5 percent at \$29.93.
The decline of the tobacco industry has been a long time coming, but it is now being accelerated by the 1998 Master Settlement Agreement.	The decline of the tobacco industry has been a long time coming, but it is now being accelerated by the 1998 Master Settlement Agreement.
That was fueling growth in cheap imports and illegal cigarette sales and outside brands made by smaller manufacturers that were partly exempted from or not complying with the 1998 Master Settlement Agreement.	That trend was fueling the growth in cheap imports and illegal cigarette sales and outside brands made by smaller manufacturers that were partly exempted from or not complying with the 1998 Master Settlement Agreement.
But the industry's growth is being eroded by the 1998 Master Settlement Agreement.	But the industry's growth is being eroded by the 1998 Master Settlement Agreement.
They have passed on to consumers through higher prices and higher contributions to their annual contributions.	They have passed on a much at the cost of their annual contributions to consumers through higher prices.
Driver Dennis Philip Morris's chief financial officer told investors yesterday US tobacco sales by volume were still recovering from the impact of the 1998 Master Settlement Agreement.	Driver Dennis Philip Morris's chief financial officer told investors Tuesday US tobacco sales by volume were still recovering from the impact of the 1998 Master Settlement Agreement.
The situation in the US tobacco business is now far more bleak than it was a year ago.	The company is pouring up to \$600 million in price promotions to try to tempt smokers back from their cheaper rivals.
Although Philip Morris' US retail shares are showing sequential improvement, it is still well below its original price target. The company is pouring up to \$600 million in price promotions to try to tempt smokers back from their cheaper rivals.	The company is pouring up to \$600 million in price promotions to try to tempt smokers back from their cheaper rivals.
The emergence has made it difficult for the big tobacco companies to push through the price increases to which they had grown accustomed.	The big tobacco companies are finding it difficult to push through the price increases they had grown accustomed to.
Markus Feldman analyst at Merrill Lynch reduced his earnings forecast for Philip Morris for next year after the price was cut.	Markus Feldman analyst at Merrill Lynch reduced his earnings forecast for Philip Morris for next year.



## Relevance-Ranking of Retrieved Documents




### Index of "Cigarette Smuggling" (by scores)


Id	Name	Downloaded Date	Relevance Score	Status
1660610	<a href="#">TOBACCO GIANT AIDING SMUGGLERS</a>	11/01/2003 08:08:00	4.88	
1656793	<a href="#">TOBACCO GIANT ACCUSED OF TURNING BLIND EYE TO SMUGGLING</a>	10/01/2003 08:17:32	4.49	
1658525	<a href="#">MEX accuse cigarette firm of smuggling dodges</a>	10/01/2003 08:17:31	3.74	
1658717	<a href="#">Smuggling of imperial tobacco cigarettes cut by half</a>	11/01/2003 08:07:59	3.66	
131012002	<a href="#">The high tax</a>	08:19:00	2.86	
11/01/2003		08:17:30	2.49	
11/01/2003		08:07:59	2.05	
10/01/2003		08:17:31	3.12	
10/12/2002		09:13:03	2.11	
23/10/2002		06:08:22	3.64	
1696756	<a href="#">Tobacco giants lose European Court bid</a>	16/01/2003 08:19:06	2.36	
1317618	<a href="#">China's Hard Habit to Break</a>	04/11/2002 07:41:52	2.37	

**1660610--TOBACCO GIANT AIDING SMUGGLERS**

Last update: 9:29:00 am **Tobacco** giant 'aiding **smugglers**'. **Imperial Tobacco** has been accused of **deliberately "turning a blind eye" to cigarette smuggling**. **Illegal** sales cost Britain £2.8 billion in lost tax last year, the Commons public accounts committee says. And Imperial's Regal and Superkings brands account for half that sum, according to Customs and Excise estimates. At one point two-thirds of exports of those brands - three billion **cigarettes** - were going to Afghanistan, Latvia, Moldova, the tax haven of Andorra and Russian enclaves of Kaliningrad, Imperial must have known that that quantity was not for local use, the committee said in a report. And Customs officers were "fobbed off" when they tried to investigate the sales, chairman Edward Leigh said. "They are not doing anything **illegal or criminal** - companies are free to export **cigarettes** where they want," Mr Leigh said. "But if you export these numbers of **cigarettes** to countries like these then surely they are turning a blind eye to **smuggling** deliberately." Imperial chief executive Sarah Davis said: "The report is based on historical data and does not reflect the high level of co-operation that exists at all levels with HM Customs Excise." Copyright: Press Association Ltd 2002, All Rights Reserved. Friday, January 10, 2003



## Sample Text: Plutonium Smuggling




**E-3083/95 by Martin Schulz (PSE) - Seizure of plutonium at Munich airport**


In the summer of 1994 a **suitcase** containing **plutonium** illegally imported into **Germany** was seized in sensational circumstances at Munich airport in the **Federal Republic of Germany**. Is The **Commission** aware of this matter and, if so, when were the **Commission** and its services, and other European agencies, informed of it? Can the **Commission** say whether the **Joint Research Centre in Karlsruhe** was involved, what services it provided for the **German** police, when it provided them, when the **plutonium** was seized, and when it was handed over to the **Joint Research Centre**?


2 -- Answer given by **Mr Papoutsis** on behalf of the **Commission** (10 January 1996)

The **Commission** would refer the Honourable Member to its earlier replies to questions about this incident (Written questions 1489/95(1) OJ C 213, 17.8. 1995] and 1508/95(2) OJ C 230, 4.9.1995] by **Mrs Breyer**. **The Commission (Euratom safeguards directorate)** was alerted by the **German** authorities in the early afternoon of 10 August, 1994, that some material might be seized. In accordance with formal agreements between the **Commission** and the **German** government this information was immediately passed by phone to the **European institute for transuranium elements (TUI)** at **Karlsruhe** to ensure that preparations were made to receive any material seized. The seizure was made by the **German** police, and the **TUI** was not involved. Its activities that night were limited to receiving the closed **suitcase** at its premises in **Karlsruhe**. Subsequently, the **TUI** performed a precise analysis of the material found inside the **suitcase**, to support the investigations carried out by Member State authorities and to determine as far as possible the source and history of the nuclear material.



### Structured Multilingual Display of Monolingual Information





#### Document Profile

**Display Language:** English  
(En, Et, De, Es, It, Et, Da, Fr, He, M, Sk)

**Title:** Seizure of plutonium at Munich airport (E-3083/95)

**Author:** Martin Schulz (PSE)

**Text Language(s):** English

**Source:** [http://ec.europa.eu/digitaljanet/res/9903/13/plutonium\\_en.html](http://ec.europa.eu/digitaljanet/res/9903/13/plutonium_en.html)

**Related Documents:** 12 ([click here to view](#))

**Retrieval Date:** 03.05.1999

**Creation Date:** 27.03.1996

**Text Length:** 287 words

**Keywords (Occurrence Frequency)**

TUI (3), Commission (7), Karlsruhe (3), seizure (6), OJ (2), plutonium (3), suitcase (3), German (4), material (4)

**Eurovoc Thesaurus Descriptors**

plutonium, import, illicit trade, Federal Republic of Germany, EAEC Joint Research Centre, airport, fraud

**Names**

**Organisations:** Commission, European Institute for Transuranium Materials (TUI), Joint Research Centre, PSE

**People:** Martin Schulz, Mrs. Breyer, Mr. Papoutsis

**Geographical References**

Germany (11) [click to view](#)

Germany (6), Karlsruhe (3), Munich (2), Germany (1), Federal Republic of Germany (1)

No Others

**Combined Nomenclature Product Groups**

**CN 2844:** "radioactive chemical elements and radioactive isotopes, incl. their fission or fissionable chemical elements and isotopes, and their compounds, mixtures and residues containing these products" (plutonium, 3)

**CN 4204:** "Trunks, suit, vanity, executive, brief, spectacle, binocular, camera, musical instrument, gun cases, holsters and similar, traveling, toilet bags, rucksacks, handbags, school satchels, shopping bags, wallets, purses, map, cigarette cases" (suitcase, 3)

**Document Summary**


**E-3083/95 by Martin Schulz (PSE)**

**Seizure of plutonium at Munich airport**


In the summer of 1994 a suitcase containing plutonium illegally imported into Germany was seized in sensational circumstances at Munich airport in the Federal Republic of Germany. The Commission (Euratom safeguards directorate) was alerted by the German authorities in the early afternoon of 10 August, 1994, that some material might be seized.

[See full text](#)

<http://www.jrc.ec.eu.int/lang/edi>



### Monolingual Keyword Identification (Indexing)




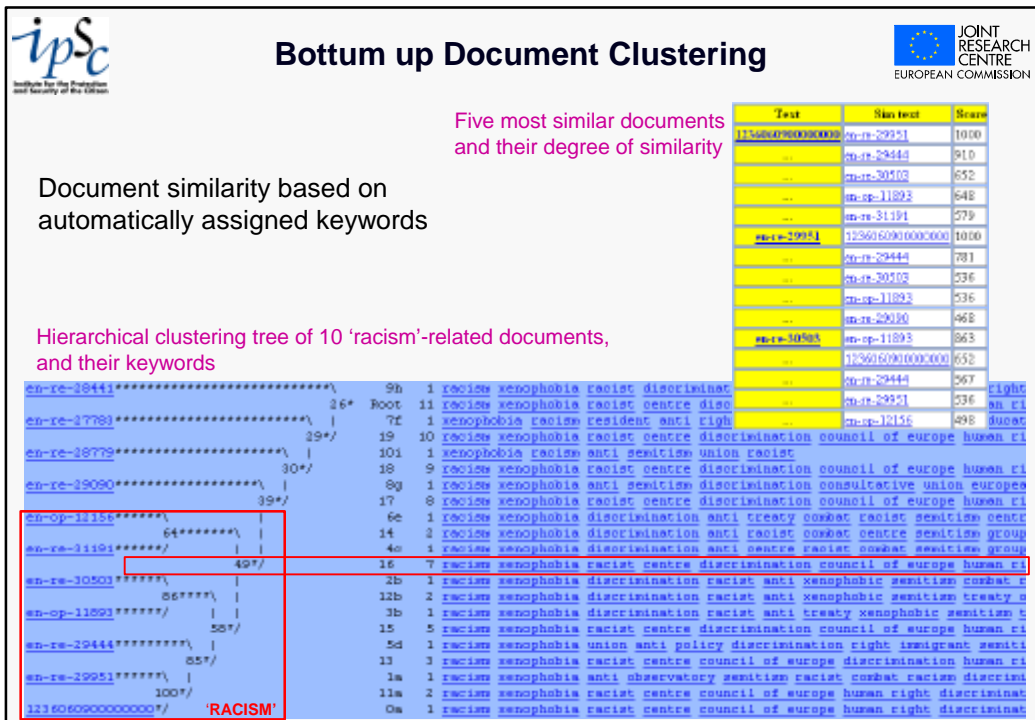
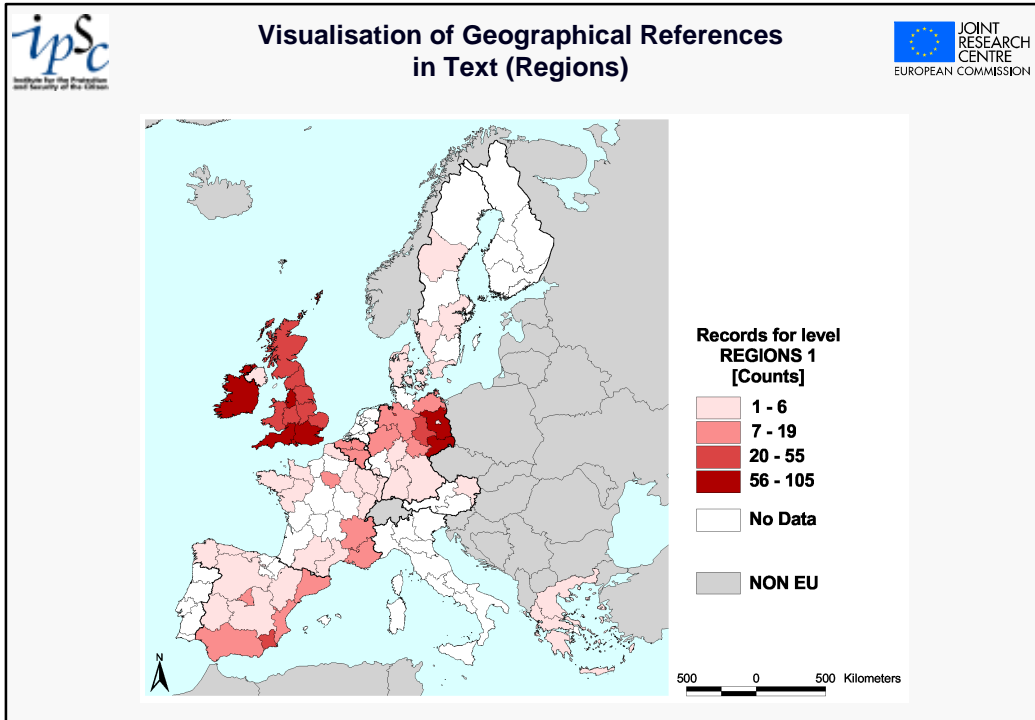
- Statistical tool identifies *statistically salient words* of the text (same language as text)
- By comparing the lemma frequency **TF** of a document with an 'expected' / average lemma frequency (reference corpus frequency **RCF**)
- Using the **log-likelihood test** (Dunning 1993). Alternatives: *chi-square*, *TF.IDF*, ...


Lemma	TF	RCF	Keyness
tui	3	5	65.26
commission	7	11231	59.81
karlsruhe	3	22	57.50
seize	4	2342	42.17
plutonium	3	437	39.94
suitcase	3	752	36.69
german	4	12738	28.69
material	4	18418	25.78
seizure	2	443	24.95
...			

**Text length:** 300 words

**Reference corpus length (BNC):** 100 million words









Institute for the Protection and Security of the Citizen

## Agenda



JOINT RESEARCH CENTRE  
EUROPEAN COMMISSION


- Introduction: Who we are and what we do
  - Joint Research Centre
  - Language Technology in IPSC
    - Document retrieval
    - Text analysis and information extraction
    - Visualisation of textual information
- **Assignment of descriptors of the multilingual thesaurus Eurovoc to texts**
  - Method; Challenges and difficulties
  - Applications:
    - Cross-lingual indexing
    - Multilingual document maps
    - Cross-lingual document similarity calculation
    - Subject-specific summarisation
- Possibilities for collaboration with the JRC



Institute for the Protection and Security of the Citizen


## Eurovoc Thesaurus

<http://europa.eu.int/celex/eurovoc>



JOINT RESEARCH CENTRE  
EUROPEAN COMMISSION

- Developed by the European Parliament (EP) and the EC's Publications Office (OPOCE), together with several national organisations
- Controlled vocabulary, wide coverage
- Multilingual (**exists in all 11 official EU languages**) !
- Eurovoc is actively used; there is a real need for automating the process
- We have access to large amounts of training material (manually indexed texts)
- Hierarchically organised into a maximum of 8 levels
  - top level: 21 fields
  - next level: 127 micro-thesauri
  - total: 5933 descriptors (version 3.0)
  - 5877 reciprocal relations (BT, NT)
  - 2730 **reciprocal associations (RT)**






## Eurovoc (Top Level and Detail)




<ul style="list-style-type: none"> <li>04 Politics</li> <li>08 International Relations</li> <li>10 European Communities</li> <li>12 Law</li> <li>16 Economics</li> <li>20 Trade</li> <li>24 Finance</li> <li>28 Social Questions</li> <li>32 Education and Competition</li> <li>36 Science</li> <li>40 Business and Competition</li> <li>44 Employment and Working Conditions</li> <li>48 Transport</li> <li>52 Environment</li> <li>56 Agriculture, Forestry and Fisheries</li> <li>60 Agri-Foodstuffs</li> <li>64 Production, Technology and Research</li> <li>66 Energy</li> <li>68 Industry</li> <li>72 Geography</li> <li>76 International Organisations</li> </ul>	<h3>28 SOCIAL QUESTIONS</h3> <ul style="list-style-type: none"> <li>2806 family</li> <li>2811 migration</li> <li>2816 demography and population</li> <li>2821 social framework</li> <li>2826 social affairs</li> <li>2831 culture and religion                             <ul style="list-style-type: none"> <li>arts</li> <li>cultural policy</li> <li>culture                                     <ul style="list-style-type: none"> <li>acculturation</li> <li>civilization</li> <li>cultural difference</li> <li>cultural identity   <ul style="list-style-type: none"> <li>RT: protection of minorities (1236)</li> <li>RT: socio-cultural group (2821)</li> </ul> </li> <li>cultural pluralism</li> <li>popular culture</li> <li>regional culture</li> </ul> </li> <li>religion</li> </ul> </li> <li>2836 social protection</li> <li>2841 health</li> <li>2846 construction and town planning</li> </ul>
--	---



## Cross-lingual Thesaurus Indexing



- Indexing, where 'keywords' are taken from a **closed list of thesaurus terms** (Eurovoc thesaurus *descriptors*).
- All descriptor terms have been translated one-to-one into all eleven official EU languages.
- Possibility of displaying descriptor terms in different languages.
- **Challenge:** Descriptor terms like DEMOGRAPHY AND POPULATION or CONSTRUCTION AND TOWN PLANNING are unlikely to occur verbatim even in texts on these issues

Spanish Text

Resolución sobre los residuos radioactivos

FUEL REPROCESSING

RADIOACTIVE WASTE

PLUTONIUM

NUCLEAR FUEL

NUCLEAR SAFETY


TRANSPORT OF DANGEROUS GOODS

WASTE DISPOSAL


STORAGE

NUCLEAR NON-PROLIFERATION

RADIOACTIVE MATERIALS



## JRC Approach – Overview




- Rule-based (linguistic) approach would be:
  - (nuclear OR radioactive) AND (accident OR leak) → NUCLEAR ACCIDENT
  - Time-consuming task
  - Rules have to be written separately for each language


vs.

- JRC's **statistical, associative approach** (bag-of-words approach)
  - Identify many (statistically or semantically) related words (*associates*) (**Training phase**)
  - Assign descriptor if many of its associates are present in text. (**Assignment phase**)

EFTA COUNTRIES	SIMPLIFICATION OF FORMALITIES
council_decision of 22 November 1993 concerning the conclusion of the Agreement in the form of an exchange of letter between the european_community and the republic_of_austria , the republic_of_finland , the republic_of_iceland , the kingdom_of_norway , the kingdom_of_sweden and the swiss_confederation relate to the amendment of the Convention of May on the simplification of formality in trade in goods	
THE council_of_the_european_union , Have regard_to_the_treaty_establish_the_european_community , and in particular Article 113 thereof , Have regard_to_the_proposal_from_the_Commission ,	
Whereas Article 11 ( 2 ) of the Convention between the european_economic_community and the republic_of_austria , the republic_of_finland , the republic_of_iceland , the kingdom_of_norway , the kingdom_of_sweden and the swiss_confederation on the simplification of formality in trade in goods ( 1 ) empower the joint_committee set_up_by_that_Convention to make recommendation for amendment to the Convention ;	
Whereas the Convention have be amend to allow for the accession of new Party ;	
Whereas the amendment in question be set_out_in_recommendation No 1/93 of the joint_committee ; whereas the Agreement in the form of an exchange of letter relate_to_that_recommendation should be approve ,	
HAVE decide_as_follow :	
Article 1: The Agreement in the form of an exchange of letter between the european_community and the republic_of_austria , the republic_of_finland , the republic_of_iceland , the kingdom_of_norway , the kingdom_of_sweden and the swiss_confederation relate_to_the_amendment_of_the_Convention of 20 May 1987 on the simplification of formality in trade in goods be hereby approve on behalf_of the Community .	
The text of the Agreement be_attach_to_this_Decision .	
Article 2: The president_of_the_council be hereby authorize to designate the person empower to sign the Agreement in_order_to_bind_the_Community .	
Do at brussels , 22 November 1993 .	



## The JRC Approach in a Nutshell (1)



EFTA COUNTRIES
SIMPLIFICATION OF FORMALITIES

council\_decision of 22 November 1993 concerning the conclusion of the Agreement in the form of an exchange of letter between the european\_community and the republic\_of\_austria , the republic\_of\_finland , the republic\_of\_iceland , the kingdom\_of\_norway , the kingdom\_of\_sweden and the swiss\_confederation relate to the amendment of the Convention of May on the simplification of formality in trade in goods

THE council\_of\_the\_european\_union , Have regard\_to\_the\_treaty\_establish\_the\_european\_community , and in particular Article 113 thereof , Have regard\_to\_the\_proposal\_from\_the\_Commission ,

Whereas Article 11 ( 2 ) of the Convention between the european\_economic\_community and the republic\_of\_austria , the republic\_of\_finland , the republic\_of\_iceland , the kingdom\_of\_norway , the kingdom\_of\_sweden and the swiss\_confederation on the simplification of formality in trade in goods ( 1 ) empower the joint\_committee set\_up\_by\_that\_Convention to make recommendation for amendment to the Convention ;

Whereas the Convention have be amend to allow for the accession of new Party ;

Whereas the amendment in question be set\_out\_in\_recommendation No 1/93 of the joint\_committee ; whereas the Agreement in the form of an exchange of letter relate\_to\_that\_recommendation should be approve ,


HAVE decide\_as\_follow :

Article 1: The Agreement in the form of an exchange of letter between the european\_community and the republic\_of\_austria , the republic\_of\_finland , the republic\_of\_iceland , the kingdom\_of\_norway , the kingdom\_of\_sweden and the swiss\_confederation relate\_to\_the\_amendment\_of\_the\_Convention of 20 May 1987 on the simplification of formality in trade in goods be hereby approve on behalf\_of the Community .


The text of the Agreement be\_attach\_to\_this\_Decision .

Article 2: The president\_of\_the\_council be hereby authorize to designate the person empower to sign the Agreement in\_order\_to\_bind\_the\_Community .

Do at brussels , 22 November 1993 .



## The JRC Approach in a Nutshell (2)



EFTA COUNTRIES ← → SIMPLIFICATION OF FORMALITIES

council\_decision of 22 November 1993 concerning the conclusion of the Agreement in the form of an exchange of letter between the european community and the republic of austria , the republic of finland , the republic of iceland , the kingdom of norway , the kingdom of sweden and the swiss confederation relate\_to the amendment of the Convention of May on the simplification\_of\_formality in trade\_in goods .

THE council\_of\_the\_european\_union , Have regard\_to\_the\_treaty\_establish the european\_community , and in particular Article 113 thereof .Have regard\_to\_the\_proposal from the Commission ,

Whereas Article 11 ( 2 ) of the Convention between the european economic community and the republic of austria , the republic of finland , the republic of iceland , the kingdom of norway , the kingdom of sweden and the swiss confederation on the simplification\_of\_formality in trade\_in goods ( 1 ) empower the joint\_committee set\_up by that Convention to make recommendation for amendment to the Convention

Whereas the Convention have be amend to allow for the accession of new Party ;


Whereas the amendment in\_question be set\_out in recommendation No 1/93 of the joint\_committee ; whereas the Agreement in\_the\_form\_of\_an\_exchange\_of\_letter relate\_to tha recommendation should be approve , HAVE decide\_as\_follow :

Article 1: The Agreement in the form of an exchange of letter between the european community and the republic of austria , the republic of finland , the republic of iceland , the kingdom of norway , the kingdom of sweden and the swiss confederation relate\_to the amendment of the Convention of 20 May 1987 on the simplification\_of\_formality in trade\_in goods be hereby approve on\_behalf\_of the Community .


The text of the Agreement be\_attach\_to this Decision .

Article 2: The president\_of\_the\_council be hereby authorize to designate the person empower to sign the Agreement in\_order\_to bind the Community .


Do at\_brussels , 22 November 1993 .




## Training: Text Normalisation



- Linguistic pre-processing = normalisation of the text
  - **Lemmatisation** (base-form reduction of words) and lower-casing:  
Transporting → transport
  - Mark-up of **multi-word expressions**  
'plant' → 'green\_plant' vs. 'power\_plant'
  - **Stop word lists** to avoid words that are not content-bearing  
general: are, they, having, in spite of, interesting,  
domain-specific: question, answer, commission, article



## Training: Produce Associate Lists



- Using a large collection of manually indexed documents (training corpus)
- For each descriptor  $D_1$ , take all documents indexed with  $D_1$
- identify the statistically salient words in each of these texts
- join these lists of statistically salient words and take the most frequently occurring words as associates. E.g. descriptor **RADIOACTIVE MATERIALS**

**radioactive**  
ukraine  
resolution  
**plutonium**  
**deuterium**  
parliament  
**nuclear**  
blottnitz  
...

+

**plutonium**  
**deuterium**  
assembly  
**nuclear**  
schmidt  
**radioactive**  
korea  
iaea  
...

+


illegal\_traffic  
chernobyl  
**radioactive**  
ukrainian  
**plutonium**  
**lithium**  
dangerous  
mox  
...

=


**radioactive (3)**  
**plutonium (3)**  
**nuclear (2)**  
**deuterium (2)**  
illegal\_traffic (1)  
chernobyl (1)  
...

- **Normalise the weight** according to a number of different criteria, e.g. apply TF.IDF formula to down-weight those lemmas that are associates to many descriptors.

➔ **Result of Training:** Weighed associate lists for all descriptors





## Associate List: RADIOACTIVE MATERIALS



➔	<b>d</b> deuterium	35.7836791092845
➔	<b>l</b> lithium	33.0805724769899
➔	<b>t</b> thorium	32.560703225522
➔	<b>t</b> tritium	32.0826451843048
➔	<b>n</b> nuclear_material	13.79399100837
➔	<b>r</b> radioactive_material	7.84970673161556
➔	<b>p</b> plutonium	6.72955494180221
➔	<b>r</b> radioactive_substance	6.43422856440347
➔	<b>n</b> nuclear	5.851612117697
➔	<b>u</b> ndine_uta_bloch_von_blottnitz	5.53278869694883
➔	<b>r</b> radioactive	4.89399300382035
➔	<b>n</b> uala_ahern	4.04706620369489
➔	<b>r</b> adon	4.03336435560442
➔	<b>m</b> ox	3.5654196472221
➔	<b>u</b> ranium	3.33954480260962
➔	<b>i</b> llegal_traffic	3.03072833135354

**Associate List: FISHERY MANAGEMENT**






	<b>fishery-related</b>	<b>fishing_resource</b>	54.4721542308385
		<b>fishing</b>	49.111563204862
		<b>fish</b>	46.196436023147
		<b>common_fishery_policy</b>	44.6741845971235
		<b>fishery</b>	44.3911518447189
		<b>fishing_activity</b>	43.3777671334009
		<b>fly_the_flag</b>	42.8744724542378
		<b>aquaculture</b>	39.2749719215554
		<b>conservation</b>	38.3480454620621
		<b>vessel</b>	37.911138722495
		<b>fishing_vessel</b>	37.8343365844963
		<b>catch</b>	36.8503034704154
		<b>fish_stock</b>	34.5289935973103
		<b>tacs</b>	34.388453583343
		<b>allowable_catch</b>	33.2880590561664
		<b>catch_quota</b>	32.2683540654092
		<b>control_system</b>	31.1753892078216
		<b>fish_for</b>	29.8386698340017
		<b>nautical_mile</b>	29.541061528168
		<b>fishing_right</b>	29.39167313160535
		<b>centimetre</b>	28.7167313160535
		<b>control_measure</b>	28.0527345432075
		<b>gross_tonnage</b>	28.0043616725124
		<b>fishing_zone</b>	27.8679836557192

**fishery-related** (red arrows pointing to rows 1-10)

**management-related** (blue arrows pointing to rows 11-20)


**Associate List: MAURITANIA**


	<b>fishery-related</b>	<b>mauritania</b>	26.5901338918222
		<b>islamic_republic_of_mauntania</b>	21.5028701117865
		<b>pole_and_line_tuna</b>	7.57561587727615
		<b>mauritanian</b>	7.38435684830453
		<b>fishery_sector</b>	6.82871829209186
		<b>cephalopod</b>	6.23513420563961
		<b>datasheet</b>	6.12232127368838
		<b>shipowner</b>	5.50033686503602
		<b>coast</b>	5.13094774515327
		<b>sea</b>	4.50387092013722
		<b>dispose_to</b>	4.36622870775623
		<b>annum</b>	4.09479477600525
		<b>inouakchott</b>	3.07236867452046
		<b>fishing</b>	3.90934015616078
		<b>fishing_zone</b>	3.93325407602887
		<b>much_oblige</b>	3.01347622659801
		<b>surface_longliner</b>	3.89812021954454
		<b>pelagic</b>	3.82550430535037
		<b>observer</b>	3.56326986434658
		<b>application_form</b>	3.50839353911244
		<b>fishery_agreement</b>	3.49778129811487
		<b>tonnage</b>	3.47894638596223
		<b>fees</b>	3.34136820139739
		<b>fish</b>	3.31932642852838

**fishery-related** (red arrows pointing to rows 1-10)

**sea-related** (blue arrows pointing to rows 11-15)




## Assignment Phase




- Pre-process new document (lemmatise, multi-word mark-up)
- Produce lemma frequency list (excluding stop words)
- Calculate similarity between lemma frequency list and descriptor associate lists, using statistical formulae

$$COSINE(d,t) = \frac{\sum_{l \in d \cap t} TFIDF_{l,d} \cdot TFIDF_{l,t}}{\sqrt{(\sum_{l \in d} TFIDF_{l,d}^2) \cdot (\sum_{l \in t} TFIDF_{l,t}^2)}}$$



## Formulae tested for descriptor assignment



$$TFIDF_{l,d} = TF_{l,d} \cdot ((\log_2 \frac{N}{DF_l}) + 1)$$

$$COSINE(d,t) = \frac{\sum_{l \in d \cap t} TFIDF_{l,d} \cdot TFIDF_{l,t}}{\sqrt{(\sum_{l \in d} TFIDF_{l,d}^2) \cdot (\sum_{l \in t} TFIDF_{l,t}^2)}}$$

$$Okapi_{t,d} = \sum_{l \in t \cap d} \log(\frac{N - DF_l}{DF_l}) \cdot \frac{TF_{l,d}}{TF_{l,d} + \frac{|d|}{M}}$$

$$SumTfidf(d,t) = \sum_{l \in d \cap t} TFIDF_{l,d} \cdot TFIDF_{l,t}$$

$$\Phi = 0.61 \frac{COSINE}{\max(COSINE)} + 0.21 \frac{Okapi}{\max(Okapi)} + 0.18 \frac{SumTfidf}{\max(SumTfidf)}$$


**Term Frequency, Inverse Document Frequency**  
 Considers occurrence frequency of lemma (l) in meta-text (TF<sub>l,t</sub>) and number of descriptors (d) for which the lemma is an associate (DF<sub>l</sub>)

**Cosine** uses TF.IDF; computes the angle of two multi-dimensional vectors (of the document (t) and of the descriptor associate list)


**Okapi** considers occurrence frequency of lemma as an associate (DF<sub>l</sub>); the number of associates in the associate list (size, |d|); the average size of descriptor associate lists (M); the total number of descriptors used (N)

**'SumTF.IDF'** adds product of TF.IDF values of associates and text lemmas

**'622'** mixed formula, uses all of the above




## 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	Inverse square Sum Tfidf <sup>2</sup>	Cosine	Rank Cosine	Okapi	Rank Okapi	Rank	Prec	Rec
F5641040706000000	FISHING CONTROLS [G]	.00144033	0.360	1	95.169	1	1	100	10
F5641020000000000	FISHING GROUNDS [H]	.00243464	0.308	2	65.018	14	2	100	20
F5641040200000000	COMMON FISHERIES POLICY [E]	.00018023	0.280	3	62.910	20	3	100	30
F5641040300000000	FISHERY MANAGEMENT [H]	.000207886	0.279	4	79.362	6	4	100	40
F5641040700000000	FISHING REGULATIONS [E]	.000197934	0.270	5	79.982	5	5	100	50
F5641040704000000	FISHING PERMIT [G]	.00306631	0.261	6	71.577	8	6	100	60
F5641040101000000	CONSERVATION OF FISH STOCKS [S]	.000189818	0.253	7	83.982	3	7	85	60
F5641040600000000	FISHING AREA [O]	.000182474	0.252	8	84.178	2	8	75	60
F5206040100000000	CONSERVATION OF RESOURCES [S]	.000234209	0.251	9	55.311	26	9	66	60
F5641050000000000	FISHERY RESOURCES	.000402863	0.232	10	75.046	7	10	60	60
F5641040800000000	CATCH OF FISH	.000313101	0.213	11	67.687	9	11	54	60
F5641040000000000	FISHERIES POLICY	.00258399	0.203	12	58.416	23	12	50	60
F5641040705000000	FISHING LICENCE	.000371136	0.181	13	57.618	25	13	46	60
F5641060100000000	FISHING FLEET	.00106478	0.179	14	63.323	19	14	42	60
F5641010000000000	FISHING INDUSTRY	.000551953	0.176	15	39.228	43	15	40	60
F5641040201000000	EUROPECHE	.000738822	0.176	16	62.240	21	16	37	60




## Evaluation of the Assignment




- Separate training and test sets
  - Train on training document set
  - Assign to test document set (ca. 600 documents)
- Compare automatic assignment with previous manual assignment
  - For each rank, calculate
    - **Precision** (correct assignments divided by all assignment up to this rank)
    - **Recall** (correct assignments up to this rank divided by n°. of man. assigned descr.)

Descriptor ID	Descriptor text	Inverse square Sum Tfidf <sup>2</sup>	Cosine	Rank Cosine	Okapi	Rank Okapi	Rank	Prec	Rec
F5641040706000000	FISHING CONTROLS [G]	.00144033	0.360	1	95.169	1	1	100	10
F5641020000000000	FISHING GROUNDS [H]	.00243464	0.308	2	65.018	14	2	100	20
F5641040200000000	COMMON FISHERIES POLICY [E]	.00018023	0.280	3	62.910	20	3	100	30
F5641040300000000	FISHERY MANAGEMENT [H]	.000207886	0.279	4	79.362	6	4	100	40
F5641040700000000	FISHING REGULATIONS [E]	.000197934	0.270	5	79.982	5	5	100	50
F5641040704000000	FISHING PERMIT [G]	.00306631	0.261	6	71.577	8	6	100	60
F5641040101000000	CONSERVATION OF FISH STOCKS [S]	.000189818	0.253	7	83.982	3	7	85	60
F5641040600000000	FISHING AREA [O]	.000182474	0.252	8	84.178	2	8	75	60
F5206040100000000	CONSERVATION OF RESOURCES [S]	.000234209	0.251	9	55.311	26	9	66	60
F5641050000000000	FISHERY RESOURCES	.000402863	0.232	10	75.046	7	10	60	60
F5641040800000000	CATCH OF FISH	.000313101	0.213	11	67.687	9	11	54	60




## Difficulty of evaluation




- BTs, NTs and RTs have to be considered.
- Number of manually assigned descriptors is small (average 5.6 per text)
- Many other descriptors are also correct.
- (Human) indexing specialists differ in their descriptor assignment (20-80% overlap).

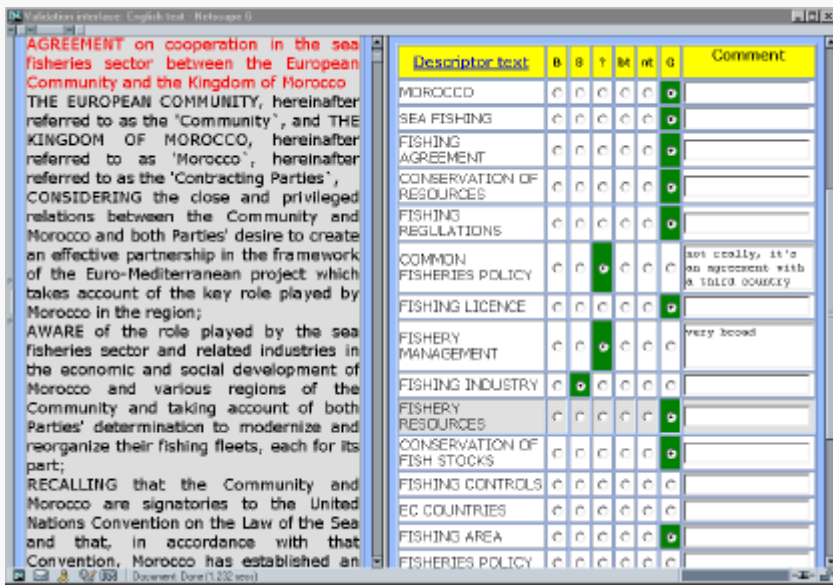
➔ **Additional evaluation** of automatically assigned descriptors by human indexer ('manual evaluation'). This provides information:

- regarding appropriateness of descriptors assigned automatically, but not manually
- on assignment overlap between two human indexers




## Manual Evaluation of the Assignment







Descriptor text	B	S	T	RT	G	Comment
MOROCCO	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
SEA FISHING	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
FISHING AGREEMENT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
CONSERVATION OF RESOURCES	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
FISHING REGULATIONS	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
COMMON FISHERIES POLICY	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	not really, it's an agreement with a third country
FISHING LICENCE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
FISHERY MANAGEMENT	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	very broad
FISHING INDUSTRY	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
FISHERY RESOURCES	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
CONSERVATION OF FISH STOCKS	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
FISHING CONTROLS	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
EC COUNTRIES	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
FISHING AREA	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
FISHERIES POLICY	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	




## Manual Evaluation – Overview




- 162 documents evaluated.
- Second evaluator reviewed previous manual assignment blindly.
- Task:
  - evaluate top ten automatic suggestions (rank 10) and
  - add missing descriptors where necessary
  - Distinguish Good, Bad, BT/NT, ?, S.
- Averages:
  - 7.5 correct descriptors per text
  - + 0.65 descriptors (BT or NT)
  - Total: 8.15 (incl. BT and NT)




- **Evaluation of previous manual assignment:**
  - 74% judged as 'Good'
  - 4% judged as 'BT' or 'NT'
  - Total: 78% agreement = benchmark for automatic assignment



## Manual Evaluation - Results

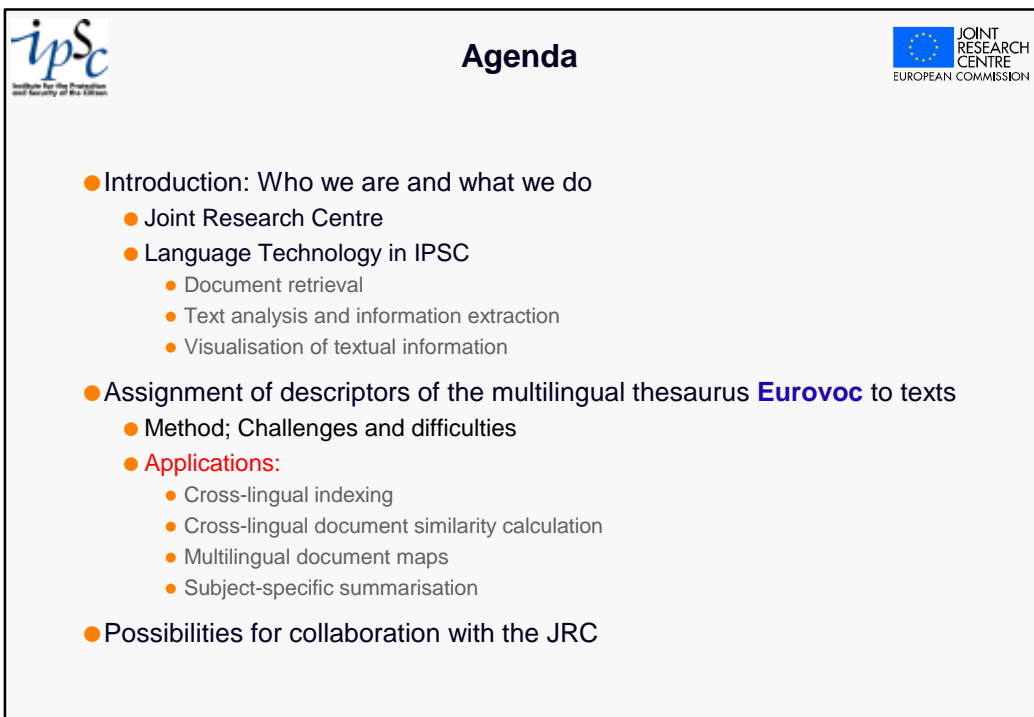
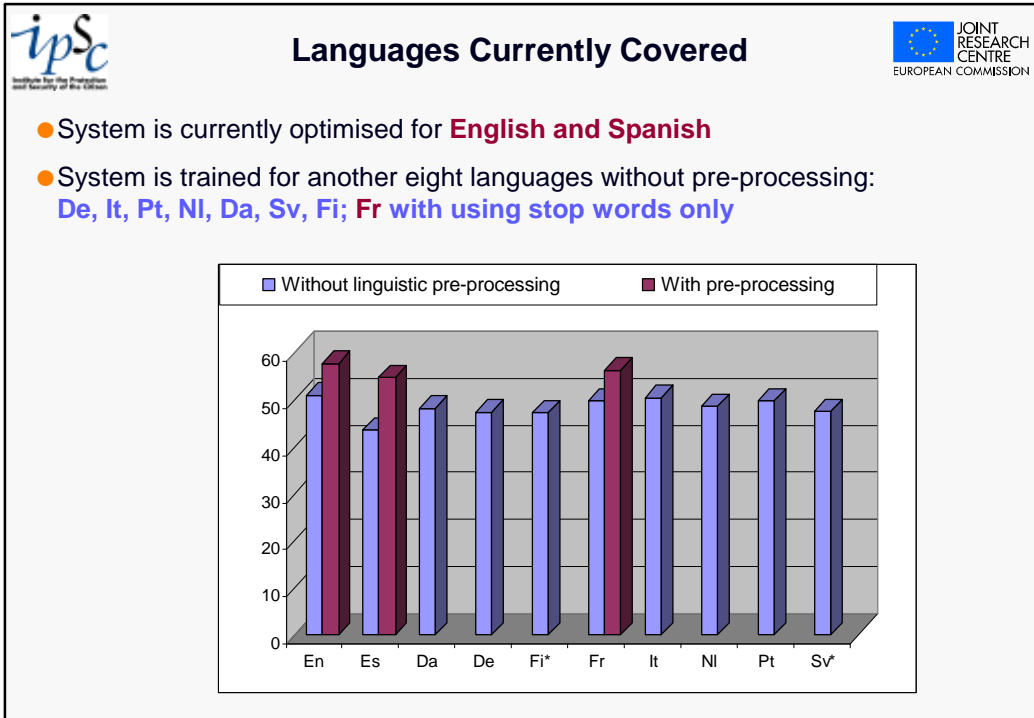


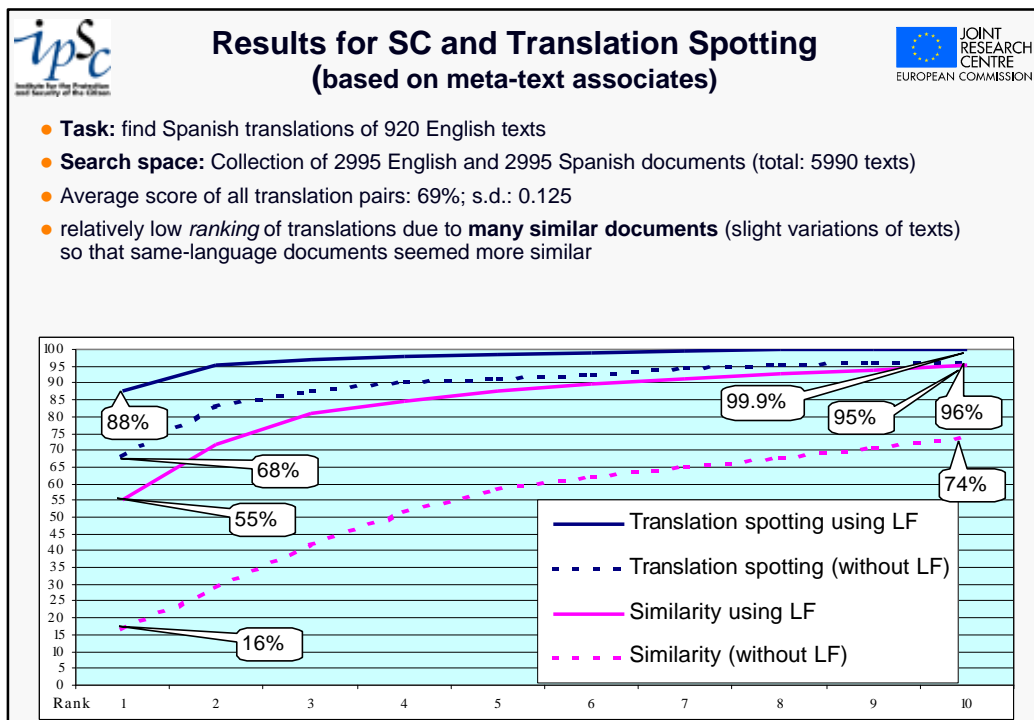
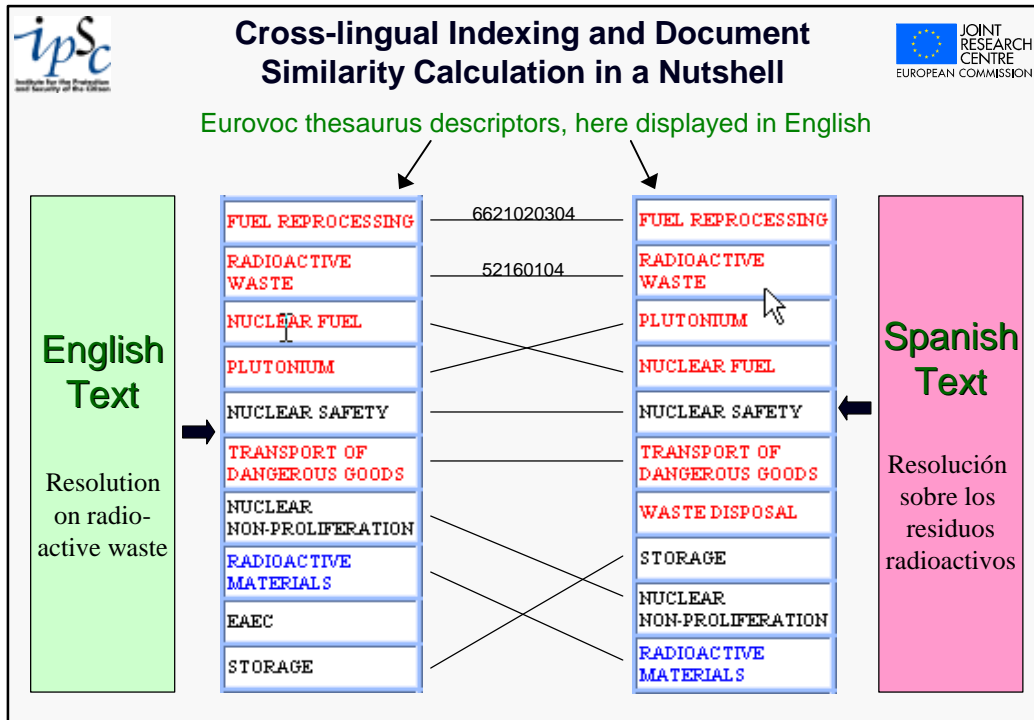
Nb of descr	Scores for exact descriptor found			Scores including BT and NT		
	P	R	F	P	R	F
1	89	12	22	94	12	22
3	78	31	45	83	31	45
5	69	46	55	75	46	57
8	60	62	61	67	63	65
10	56	71	62	62	72	66




- Correct descriptors compared to benchmark of manual assignment (78% G + BT + NT):
 


$67 / 78 = 86\%$
- **Open questions:**
  - What about the **33% incorrect ones** (B + S + ?)
  - Where to find the **37% missing** descriptors?
  - How many descriptors to present?
  - How to avoid BT-NT co-occurrence?







## Document Map (Cartia's ThemeScape)



ThemeScape Map Viewer: JRC Full Text - Microsoft Internet Explorer

Address: http://demo.cartia.com/subjectmap1124.html

Map Layers: Search: Topic List: Flags:


Search for the map topics you would like to display:


Topic	Count
Fraud	76
Projects	66
Production	53
Outline	47
Services	46
Regions	40
Irregularities	37
Health	36
Committee	35
Env	35
Customs	32
Cost	30
Expert	24
Transport	23
Meas	22
Expenditure	22
Transit	22

Search Options


Look for: all of the selected topics

Limit results to: 100 documents





## Subject-specific Summarisation



Aspect: 'Nuclear Accident'


Text title: "Resolution on the 10th anniversary of the Chernobyl accident"

Eurovoc descriptors:


NUCLEAR ACCIDENT
UKRAINE
NUCLEAR SAFETY
NUCLEAR POWER STATION
DECOMMISSIONING OF POWER STATIONS

The resolution adopted by the Commission on 8 April 1987 on the consequence of the Chernobyl accident and on the outline communication from the Commission of the European Communities to the Council on the consequence of the Chernobyl accident ( COM(86)0327 ) and the communication from the Commission of the European Communities to the Council on community action to be taken in response to the Chernobyl accident ( COM(86)0276 ) [ OJ C 125 , 11.5.1987 , p. 96 ] - (...)

1. Emphasize that 10 year after the Chernobyl nuclear disaster , which thousand\_of death by radioactive contamination and which be still have tragic consequence for the health of million of people particularly as a result of the increase in the incidence of cancer and leukaemia and for the state of the environment , the risk of a fresh accident remain , both on the Chernobyl site itself and in all other nuclear power station , (...)



Resolution on the 10th anniversary of the Chernobyl accident  
**Aspect: 'Ukraine'**



JOINT RESEARCH CENTRE  
EUROPEAN COMMISSION

**Eurovoc descriptors:**

- NUCLEAR ACCIDENT
- UKRAINE
- NUCLEAR SAFETY
- NUCLEAR POWER STATION
- DECOMMISSIONING OF POWER STATIONS

the european union...

have regard to its resolution of 15 December 1993 on the proposal for a Council Directive (1993/104/EEC) on the protection of the environment for the period of efficiency and safety of nuclear power stations in certain non-member countries (COM(93)0477) [ OJ C 125 , 11.5.1997 , p. 92 ]...

have regard to its resolution of 15 December 1993 on nuclear safety in the country of eastern Europe and the Commonwealth of Independent States (CIS) [ OJ C 292 , 24.1.1994 , p. 207 ]...


have regard to its resolution of 13 December 1993 on nuclear safety in the country of eastern Europe and the Commonwealth of Independent States, which thousand of deaths by radioactive contamination and which will have tragic consequences for the health of million of people particularly as a result of the increase in the incidence of cancer and leukaemia and for the state of the environment, the risk of a fresh accident remain, both on the Chernobyl site itself and in all other nuclear power station...

is aware that the Chernobyl nuclear power plant contribute about 17 to the electricity supply of Ukraine...


is aware that the Chernobyl nuclear power plant contribute about 7 to the electricity supply of Ukraine...

point out that the Ukrainian Government admit that Ukraine be one of the much energy-intensive country in the world, use at least 3 time the amount of energy per unit of GNP as the European Union, (...)

call on the commission to continue the Chernobyl project, which provide medical aid to Ukraine, Belarus and Russia, and request that aid to victim of the disaster be reinforce, as well as support for NGOs participate in this aid...



**Agenda**



JOINT RESEARCH CENTRE  
EUROPEAN COMMISSION

- Introduction: Who we are and what we do
  - Joint Research Centre
  - Language Technology in IPSC
    - Document retrieval
    - Text analysis and information extraction
    - Visualisation of textual information
- Assignment of descriptors of the multilingual thesaurus Eurovoc to texts
  - Method; Challenges and difficulties
  - Applications:
    - Cross-lingual indexing
    - Cross-lingual document similarity calculation
    - Multilingual document maps
    - Subject-specific summarisation
- Possibilities for collaboration with the JRC



## Collaborating with the JRC



- We are happy to enter a **collaboration**
- M.Sc. or Ph.D. **thesis** subjects
- **Informal exchange**, common experiments
- We can participate in **EU-funded projects**
- Occasionally, there are possibilities for **post-doc grants**
- **Internships** of a minimum duration of 2 months (currently unpaid; soon ca. 700 €. Life is cheap in Ispra)
- See <http://www.jrc.it/langtech> for more information