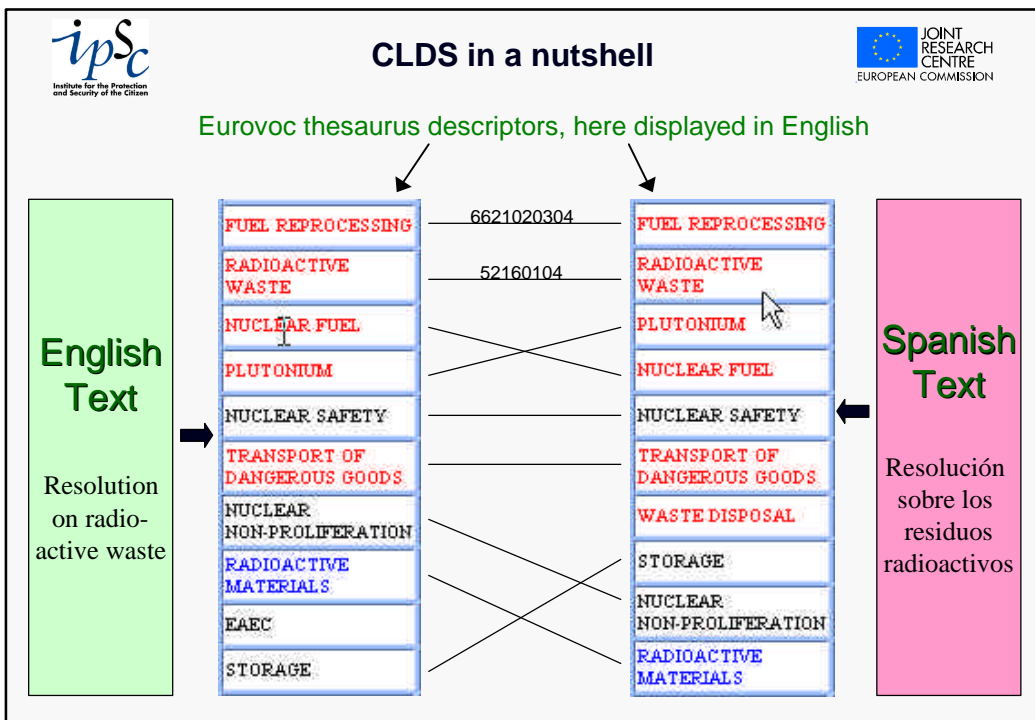


## Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC

Third International Conference on Intelligent Text Processing and Computational Linguistics  
**CICLing-2002** (<http://www.cicling.org>), Mexico City, Mexico, 17-23 February 2002

**Ralf Steinberger & Bruno Pouliquen & Johan Hagman**  
European Commission – Joint Research Centre (JRC)  
Institute for the Protection and Security of the Citizen (IPSC)  
Cyber-security and New Technologies for Combating Fraud (CSCF)

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



- What is (cross-lingual) document similarity (DS)?  
Why and how to calculate it?  
How to measure it?
- Document similarity calculation (DSC) in the context of the JRC activities
- Assigning Eurovoc thesaurus *descriptors* to text
- Using the result to calculate document similarity
- Implementation details
- Limitations of this method
- Planned work

- Intuitively clear, but it is hard to put one's finger on it.
- Could be based on:
  - degree of lexical overlap between two texts
  - map document contents onto a complex conceptual space (LSI, Landauer & Littman 1999)
  - stylistic information (sentence length, type-token ratio, word variation, ...)
  - meta-information (document type, format, author name, source, time of writing, ...)
- Our approach:
  - map texts onto an existing knowledge structure (multilingual Eurovoc thesaurus), using statistical methods
  - Eurovoc acts as a **multilingual conceptual interlingua**
  - requires thesaurus and training material (manually indexed texts)
  - does not require additional language pair-specific data (dictionaries)



## Motivation to calculate document similarity




- produce a ranked list of documents that are similar to a given one, even if they are written in different languages
- allow navigation through a multilingual document collection, using document maps (Hagman et al. 2000, Steinberger et al. 2000)
- automatic multilingual clustering or classification of documents
- compile a collection of parallel texts found on the internet (Resnik 1999, Smith, 2001)




## How to measure and evaluate document similarity calculation




- Non-trivial question: how to judge (even intuitively) document similarity
- Questions:
  - 3-page text vs. its 20-line summary of this text vs. related 3-page text
  - Should text length play a role?
  - Should document language be a factor?
- Intuitively, the *translation* of a text should be most similar (< 100%)
- Evaluation is difficult. Non-optimal solution:
  - success rate of spotting translations of a text as an evaluation criterion
  - but: tasks of translation spotting and similarity calculation are not the same:
    - search space for translations is restricted to other languages
    - text length is clearly a factor
    - system can be optimised for translation spotting.  
What does this say about similarity calculation? => two separate experiments




## DSC in the context of the JRC's activities



- **The JRC**
  - is a Directorate General (DG) of the EC ([www.jrc.it](http://www.jrc.it) ; [www.jrc.cec.eu.int](http://www.jrc.cec.eu.int))
  - Employs ca. 2500 people in 8 institutes in 5 locations (I, E, D, NL, B)
- **Goal of the LT group:** put together a modular system with three main components (IDoRA system; <http://www.jrc.it/langtech>):
  - **Retrieval** of potentially relevant texts (from the internet, etc.) in a variety of languages, using agent technology (Scheer et al. 2000)
  - **Analysis:** Extraction of a variety of information aspects from these texts
    - recognise **key words**, subject domains and language of texts, references to geographical places, to people, to products, etc.
    - Calculation of the **similarity of documents**; clustering and classification of texts
  - **Visualisation** of the contents
    - of individual documents in *document profiles*
    - of whole text collections in **document maps** (Hagman et al. 2000, Steinberger et al. 2000)
- Small team; many languages to deal with => **usage of mainly statistical methods**



## Document Profile



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

**Retrieval Date** 03.05.1999

**Creation Date** 27.03.1996

**Language(s)** English (97% probability)

**Source** [http://cnnh.com/digitaljam/wires/9906/13/plutonium\\_eu.html](http://cnnh.com/digitaljam/wires/9906/13/plutonium_eu.html)

**Display Language** English (En, Fr, De, Es, It, Pt, Da, Fi, He, Nl, Sv)

Free Indexing Terms

TUI, Commission, Karlsruhe, seizure, OJ, plutonium, suitcase, German, material

Eurovoc Indexing Terms

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

Names

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

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

Combined Nomenclature Product Groups

**CN 2844:** "radioactive chemical elements and radioactive isotopes, incl. their fissile or fertile 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; travelling, toilet bags, rucksacks, handbags, school satchels, shopping-bags, wallets, purses, map, cigarette cases" (**suitcase**, 3)

Document Summary

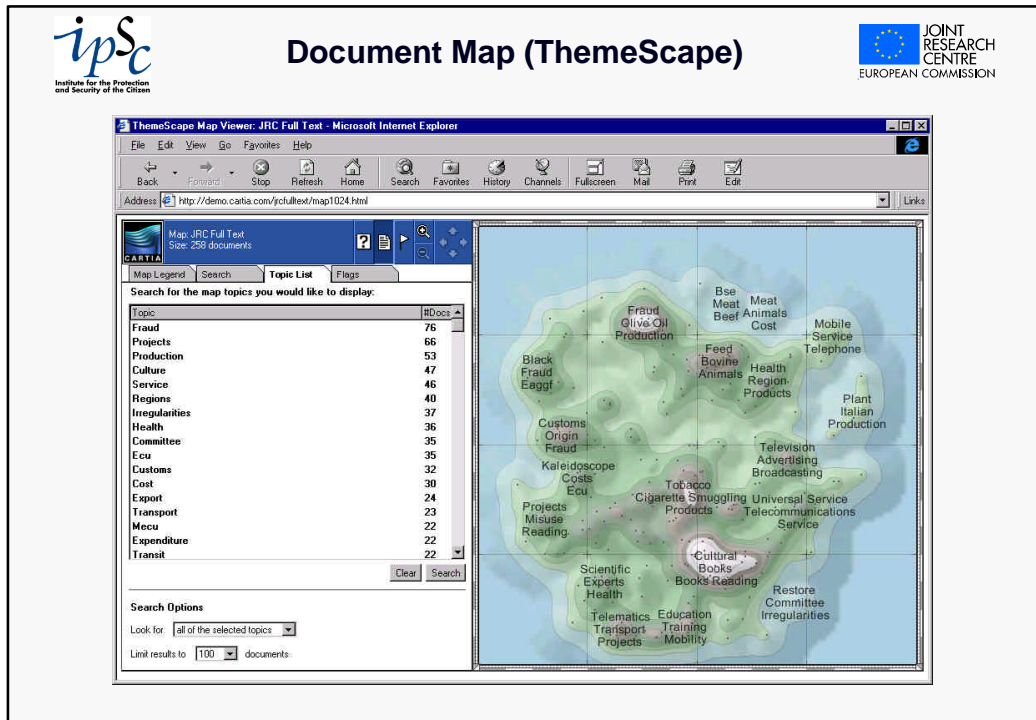
**E-3083/95 by Martin Schulz (PSE)**

**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.

Geographical Profile

<b>Relevance:</b>	70%
<b>Germany</b>	100%
Germany, German, München, Karlsruhe	
<b>Others:</b>	0%



**Assignment of Eurovoc Descriptors**

- Eurovoc thesaurus
- Assignment using statistical methods
  - Training phase, using collection of manually indexed texts
  - Assignment phase



## Eurovoc Thesaurus

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




- 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**) !
- 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><b>28 Social Questions</b></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>	<p><b>28 SOCIAL QUESTIONS</b></p> <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><b>2831 culture and religion</b></li> <li>arts</li> <li>cultural policy</li> <li><b>culture</b></li> <li>acculturation</li> <li>civilization</li> <li>cultural difference</li> <li><b>cultural identity</b></li> <li>RT: protection of minorities (1236)</li> <li>RT: socio-cultural group (2821)</li> <li>cultural pluralism</li> <li>popular culture</li> <li>regional culture</li> <li>religion</li> <li>2836 social protection</li> <li>2841 health</li> <li>2846 construction and town planning</li> </ul>
---	---




### Assigning Eurovoc thesaurus descriptors to texts




- **Challenge:** Descriptor terms like DEMOGRAPHY AND POPULATION OR CONSTRUCTION AND TOWN PLANNING are unlikely to occur as such even in texts on these issues
- **Training phase:** we generate, for each descriptor, large lists of associated lemmas (*associates*)
- Example descriptor 56410401: **FISHERY MANAGEMENT**
- **Assignment phase:** the more associates for a descriptor we find in a text, the more likely it is that this descriptor is appropriate

Lemma	TFIDF	DF (of lemma)	Log-likelihood
f fishery	1913.83	117	1382
f fish	1369.68	103	873
f conservation	1174.88	102	828
f fishing	1162.04	79	816
f stock	985.59	72	701
f management	1252.94	137	602
f fish_stock	750.92	33	589
f vessel	1119.39	114	585
f organization	1283.43	124	549
f fishery_management	516.04	9	445
f migratory_fish_stock	587.17	9	428
f subregional	611.74	23	423
f mediterranean	644.84	89	383
f fishery_resource	505.12	38	360
f flag	598.78	67	357




### Training: produce lists of associates




- Minimal linguistic text **pre-processing**
  - lemmatisation (base form reduction of words)
  - mark-up of multi-word units (MWU), identified with 'quick and dirty' procedure:
    - produce frequency list of all groups of two, three, etc. lemmas from training corpus
    - do not allow stop words as first or last word
    - hand-select reasonable terms (ca. 50%) Of the table of contents in
    - extensive stop word lists (lemmas)
- Combine all texts manually indexed with one descriptor into a **meta-text**.
- **Compare** the meta-text **lemma frequency** list with the lemma frequency list of the whole training corpus (reference corpus), using the **log-likelihood test** (Dunning 1993). Alternatives: *chi-square*, *TF.IDF*, ... (Kilgarriff 1996)
- **Result:** a **list of keywords** (associates) + their **keyness** (weight; relevance for the contents of the meta-text)
 

word	freq	freq	keyness
commission	1,486	18,362	9,896.1
member	518	15,940	2,605.5
eec	146	151	1,610.9
- **Apply TF.IDF** formula to down-weight those lemmas that are associates to many descriptors.




## Assignment phase




- for a new text, pre-process and compare the lemma frequency list of this new text with all descriptor associate lists by
- using a statistical algorithm to identify those descriptor associate lists that are most similar to the text's lemma list → most appropriate descriptors
  - **Cosine** (Salton 1989): **best precision for single formula**, compared to manually assigned descriptors
  - **Okapi** (Robertson et al. 1994) **best** when used as input **for DSC** !
  - **Sum TF.IDF**
  - **mixed algorithms** (e.g. '622') **best precision (cosine + 5%)**, but computationally heavier and harder to use for DSC
  - ...

→ **Result:** a ranked list of the most suitable descriptors for this text



## Formulae tested for descriptor assignment



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

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

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

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

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


**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




### Sample assignment result (cosine formula)




**Title: Resolution on radioactive waste** (6 manually assigned descriptors)

**Cosine** ranks: 1, 2, 3, 4, 8, 19 (52060101 - waste disposal)  
**Okapi** ranks: 1, 2, 3, 4, 6, 11  
**622** ranks: 1, 2, 3, 4, 7, 11  
**Sum TF.IDF** ranks: 5, 9, 11, 13, 16, 22

Descriptor ID	Descriptor text	Inverse square Sum Tfidf <sup>2</sup>	Cosine	Sum TFIDF	Okapi	622	Rank	Prec	Rec
6621020304000000	FUEL REPROCESSING	.000988382	0.717	161875	100.25	0.697	1	100	16
6621010101000000	PLUTONIUM	.000943596	0.661	156494	60.84	0.649	2	100	33
5216010400000000	RADIOACTIVE WASTE	.00057321	0.518	201685	88.70	0.577	3	100	50
6621010200000000	NUCLEAR FUEL	.000693454	0.511	164644	73.84	0.565	4	100	66
6621010100000000	RADIOACTIVE MATERIALS	.000735633	0.314	95401	47.73	0.428	5	80	66
6621020200000000	NUCLEAR SAFETY	.000222327	0.285	286289	55.93	0.427	6	66	66
0816040102000000	NUCLEAR NON-PROLIFERATION	.000439714	0.244	123791	52.91	0.392	7	57	66
4811030902000000	TRANSPORT OF DANGEROUS GOODS	.000219582	0.240	244148	53.79	0.397	8	62	83
7606030200000000	IAEA	.000956492	0.228	53192	21.30	0.346	9	55	83
1006010300000000	EAEC	.000486021	0.198	90858	47.02	0.357	10	50	83

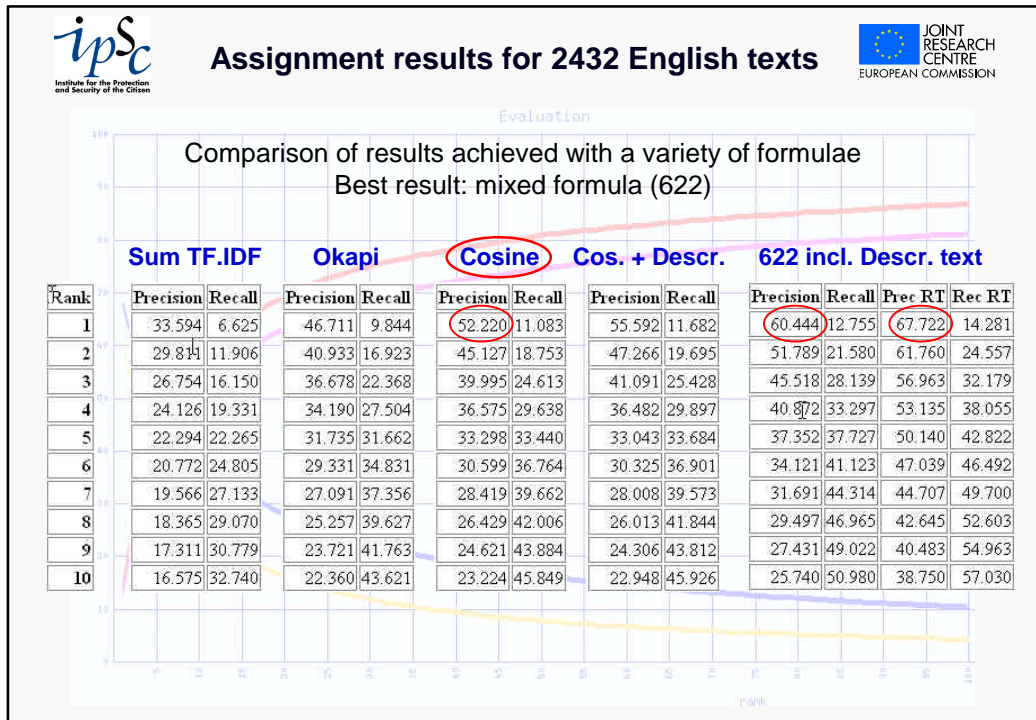


### Evaluation of the assignment



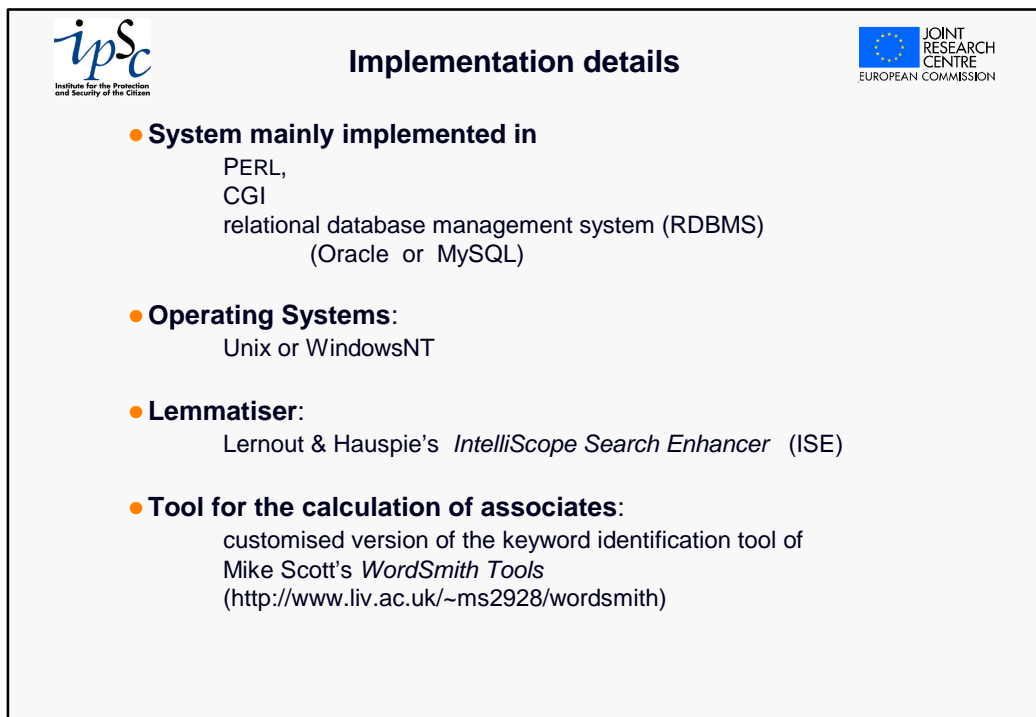
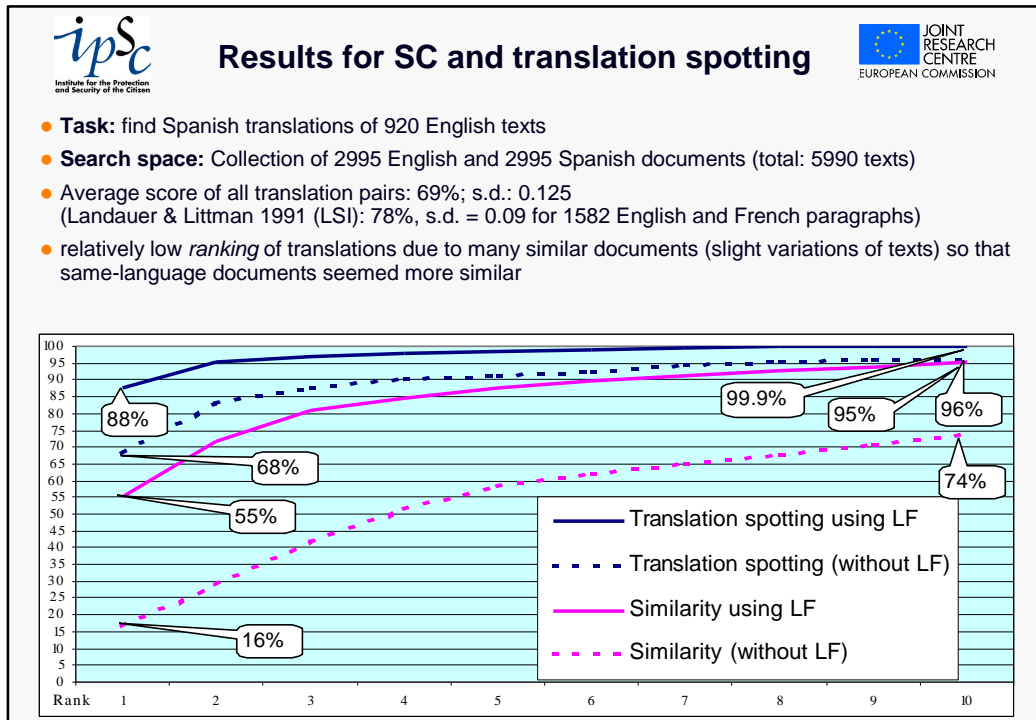
- Comparison with manually assigned descriptors
- Manually assigned descriptors vary between 1 and over 20
- Precision and recall can be calculated for each rank
- Average number of descriptors was 6.59 in EP training collection and 5.21 in OPOCE test collection
- Evaluation should consider **hierarchical structure** (*broader terms* BT and *narrower terms* NT), as well as cross-hierarchy relations (*related terms* RT)
- Human indexers are advised not to use both BT and NT, while our system exclusively follows the similarity measure and will assign both
- Human indexers do not assign same descriptors and same number of descriptors either, and their assignment is inconsistent (day to day, over the years, etc.); past studies: **only 30% to 80% overlap between human indexers**

→ The manual results are not an absolute criterion for the assignment quality.



**(Cross-language) Document similarity calculation**

- **Input:** long lists of automatically assigned Eurovoc descriptors and their relevance value (vectors)
- **Calculation** of the similarity between these vectors using the cosine formula (best result for *descriptor assignment* produced with Okapi formula)
- **Evaluation** by checking whether the translation of the input text is identified as the most similar document (translation spotting)
  - but: performance for translation spotting can be optimised (document length; search space restricted to the other language)
- ➔ Two separate experiments:
  - general **similarity calculation** (no length restriction, full search space)
  - **translation spotting** (length restriction; search only texts of other language)



## Limitations of this Method

- Performance depends heavily on the quantity and quality of the training data
  - not enough training data (received from EP) for all descriptors
  - very uneven distribution (descriptors used between 0 and 1000 times)
  - EP uses specific sublanguage (mainly legal, formal)
  - we expect better results when adding OPOCE training data
- Eurovoc has wide coverage, but is not very detailed
  - when mapping document contents onto this coarse knowledge structure we lose some information (e.g. highly scientific texts)
- Current system is restricted to the 11 official EU languages

## Planned work

- Experiment: train system on texts gathered automatically from the internet to add more associates and to apply to more languages
- Improve performance
  - add the OPOCE training data (different source, different text types, more!)
  - better text normalisation (multi-word term mark-up, stop word lists, etc.)
  - better cleaning of the training collection, e.g.
    - take out foreign language parts in documents
    - identify irregular texts (e.g. original has annex, translation does not, etc.)
  - experiment more with various parameters and new formulae
- Apply to real world scenarios
  - apply to more languages
  - incorporate with a working system at customer sites
- Follow up any suggestions you might have