



## Providing Cross-lingual Information Access in Multilingual Text Collections

JRC lecture organised by the Scientific Committee of the IPSC

11 April 2002

Ralf Steinberger

European Commission – Joint Research Centre (JRC)  
Institute for the Protection and Security of the Citizen (IPSC)  
Cyber-Security and New Technologies for Combating Fraud Unit (CSCF)  
Anti-fraud Information Management Sector (AIM)

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



### The 'Team'

Ralf Steinberger

Bruno Pouliquen

(Johan Hagman)

(Stefan Scheer)

(Marco Palazzini)

(Giovanni Valerio)



## Introduction - Terminology

- **Computational Linguistics (CL)**  
*is a **discipline** dealing with computers and natural language*
- **Natural Language Processing (NLP),  
Language Technology (LT),  
Language Engineering (LE)**  
  
*are terms referring to the application areas of CL.*
- (Computational) Linguistics is **not an exact science!**  
*Language is closely related to cognition.*

## Introduction - Application Areas

- **Machine Translation (MT)**
- **Spell checkers**
- **Grammar / syntax / style checkers**
- **Speech recognition and synthesis:**  
*to make machines more user-friendly or for dictation machines, etc.*
- **Document / text retrieval:** *find one or more documents matching certain search criteria (e.g. a certain letter or texts on a certain subject) in a large repository of documents*
- **Automatic text summarisation / abridgement:** *present only the most relevant information of a text*
- **Document filtering / Personal information filtering:** *select (user-) relevant documents from a large flow of incoming messages such as email messages or newswires and present only these to the user*
- **Document classification:** *assign documents to categories of a given subject classification, e.g. decide whether newswires relate to sports, finance, agriculture, etc.*

## Introduction - Application Areas (2)

- **Automatic keyword assignment to texts (indexing):**  
assign keywords automatically to a text to facilitate retrieval (e.g. for libraries)
- **Information Extraction:** e.g. identifying proper names, company names, place names, date expressions, etc.; extraction of whole scenarios from texts (e.g. succession of company leaders, bank transfers, etc.)
- **Computer-Assisted Language Learning (CALL)**
- **Monolingual or multilingual document generation**
- **Translator's workbench:** a set of tools improving translators' efficiency including MT, translation memory, on-line dictionaries (bilingual, synonyms, ...), automatic terminology extraction, a terminology bank, etc.
- **Document navigation:** in a large repository of documents, find relevant ones and move from one document to other related ones, e.g. by comparing the overlap of common indexing terms or using other similarity indicators
- ...

## Question

### Why are results of Machine Translation and other software not better?

Answer 1: There are many exceptions in language;  
there are interfering and soft rules;  
language permanently changes, etc. ...

Answer 2: Language contains little information.  
The rest of the information is in our heads ...

## Knowledge Contained in Language

- **Linguistic knowledge:** knowledge contained in the language

She **wrote** a letter to her parents at the bar.

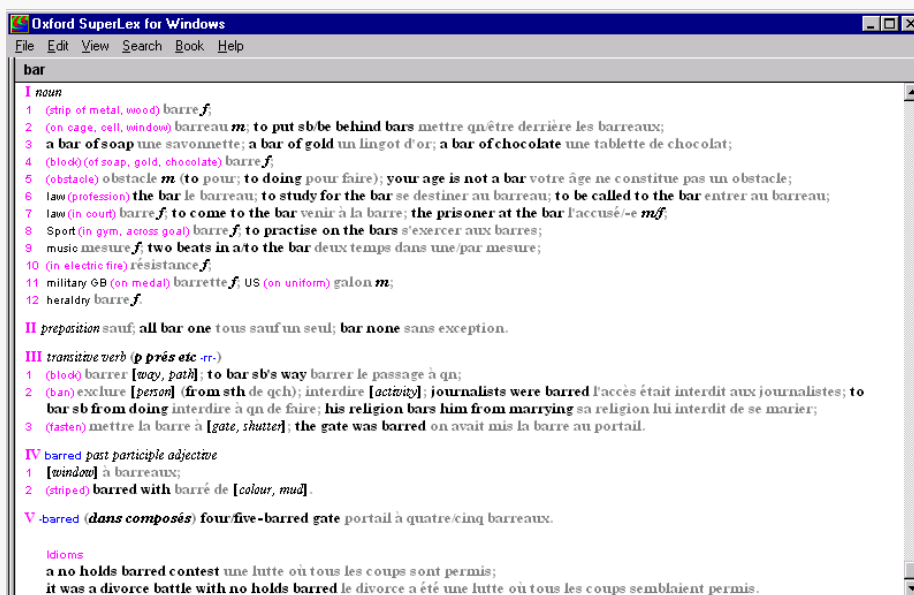
- past tense (tense system of the grammar; morphology)
- SU + write + DOBJ + IOBJ (+ ADV)  
(sub-categorisation information of the verb *write*)
- SU ≠ OBJ (syntactic word order knowledge)

- **Contextual knowledge:** knowledge contained in the same text  
(or present due to common environment)

... . **She** wrote a letter to **her** parents sitting at the **bar**.

- Anaphora (pronoun) resolution
- Word sense disambiguation ('**bar**' = 1. place for drinks, 2. straight piece of metal, 3. legal organisation, 4. stripe, ...)

## En-Fr Dictionary Entry for 'bar'



**bar**

**I noun**

- (strip of metal, wood) barre *f*;
- (on cage, cell, window) barreau *m*; to put sb/be behind bars mettre qn/être derrière les barreaux;
- a bar of soap une savonnette; a bar of gold un lingot d'or; a bar of chocolate une tablette de chocolat;
- (block) (of soap, gold, chocolate) barre *f*;
- (obstacle) obstacle *m* (to pour; to doing pour faire); your age is not a bar votre âge ne constitue pas un obstacle;
- law (profession) the bar le barreau; to study for the bar se destiner au barreau; to be called to the bar entrer au barreau;
- law (in court) barre *f*; to come to the bar venir à la barre; the prisoner at the bar l'accusé/-e *m/f*;
- sport (in gym, across goal) barre *f*; to practise on the bars s'exercer aux barres;
- music measure *f*; two beats in a to the bar deux temps dans une par mesure;
- (in electric fire) résistance *f*;
- military GB (on medal) barrette *f*; US (on uniform) galon *m*;
- heraldry barre *f*.

**II preposition** sauf; all bar one tous sauf un seul; bar none sans exception.

**III transitive verb** (p prés etc -rr-)

- (block) barrer [way, path]; to bar sb's way barrer le passage à qn;
- (ban) exclude [person] (from sth de qch); interdire [activity]; journalists were barred l'accès était interdit aux journalistes; to bar sb from doing interdire à qn de faire; his religion bars him from marrying sa religion lui interdit de se marier;
- (fasten) mettre la barre à [gate, shutter]; the gate was barred on avait mis la barre au portail.

**IV barred** past participle adjective

- [window] à barreaux;
- (striped) barred with barré de [colour, mud].

**V -barred** (dans composés) four five-barred gate portail à quatre/cinq barreaux.

**Idioms**

a no holds barred contest une lutte où tous les coups sont permis;  
it was a divorce battle with no holds barred le divorce a été une lutte où tous les coups semblaient permis.

## Knowledge Contained in Language (2)

- **World knowledge:** knowledge not contained in the language or in the text such as knowledge on the shape and character of things in the world, knowledge on inference mechanisms, etc.
  - **The man saw the elephant with the telescope.**  
(telescope is a tool to see)  
  
**The elephant saw the man with the telescope.**  
(telescope is an attribute of the man)
  - A **small** elephant vs. A **large** fly
- ➔ Language Technology applications are mostly restricted to the information contained in the texts.
- ➔ Teaching computers to use contextual knowledge and world knowledge, inference mechanisms, etc. is very difficult and labour-intensive

## Common Problems in Language Engineering

- **Ambiguity of POS:** e.g. saw (Noun or Verb? see or saw?)  
→ need for POS disambiguation and sometimes semantic disambiguation
- **Structural (syntactic) ambiguity**  
 $[Time]_{SU} flies_V [like\ an\ arrow]_{ADV}$  vs.  
 $[Time\ flies]_{SU} like_V [an\ arrow]_{OBJ}$
- **Semantic ambiguity:** e.g. bar, suit
- **Ellipsis:** Es:  $\emptyset$  cantaría (1<sup>st</sup>, 3<sup>rd</sup>) = En: */he/she* would sing
- **Morphological ambiguity:** Es: cantaría = It: canterei (1<sup>st</sup>) / canterebbe (3<sup>rd</sup>)
- **Multi-word expressions:** (power) plant vs. 'time flies',  
En: colour liquid crystal display vs. De: Flüssigkristallfarbbildschirm)

## Common Problems in Language Engineering (2)

- **Orthographic errors in texts** (typos, bad OCR output)
- **Discontinuous elements:**  
*Er brachte das Buch 1993 heraus. (herausbrachte = published)*  
*make good use of*
- **Unclear sentence borders:** (... in the U.S.A. [Chancellor Kohl promised 3.2 million DM to Mr. Smith while Mr. Chirac avoided the issue.] ... )
- **Structural difference between languages** (complex MT transfer)  
*He swam across the river. ≈*  
*Il a traversé la rivière en nageant.*
- Use of metaphors, stylistic and lexical variation, ...
- ...

## Approaches in Language Engineering

- **Rule-based (symbolic, linguistic) methods**
  - description of all the rules in a formalism, having large and completely coded dictionaries → complex applications are very time-consuming to develop.
  - typical applications are morphological analysers, tools to recognise names and multi-word expressions, MT and CALL systems
- **Statistical methods**
  - using absolute or relative frequency of words, letters, co-occurrence of words, etc.
  - typical applications are language recognisers based on bigram frequency and summarising / abridging systems.
  - systems are often trained on manually encoded texts (e.g. POS taggers)
- **Hybrid methods, combining rules and statistics**
  - e.g. POS tagging, lemmatisation, etc. for applications is done in a rule-based manner and word frequency statistics are used to identify indexing terms or similarities between documents, or
  - rule-based MT system identifies syntactic or lexical ambiguity and statistical data is used to choose the most likely reading



## What do we do? Why? For whom?



### ● What?

- Develop, evaluate, purchase and integrate linguistic software tools that help users 'deal' with textual information (find, analyse, display).

### ● Why?

- Help users to turn the phenomenon 'information overflow' into a rich and usable source of knowledge
- Help users to overcome the 'language barrier', which is especially important in an international organisation such as the EC

### ● For whom?

- Main user currently is the EC's anti-fraud organisation OLAF; also others
- Techniques are useful in any other field and for any other organisation, private or public

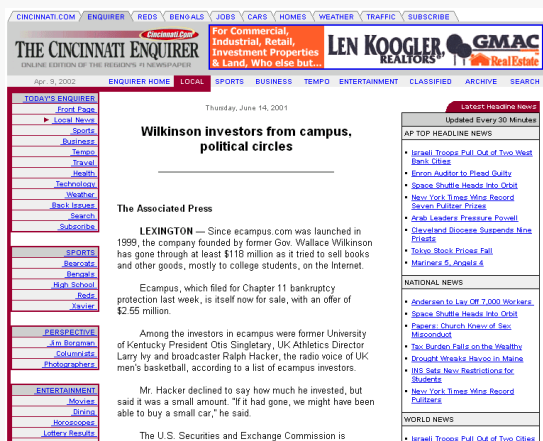


## Goal of JRC's Language Technology work




- **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** (Steinberger et al. 2002); 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


- OSILIA Project (Open Source Intelligence Library for Internet Abuse); JRC-internal collaboration with *Cyber-Security* and *Web Technology* sectors
- Automatic, daily search of ca. 20 online newspapers and other news sites written in English and German for articles about a certain subject field (e.g. 'internet abuse')
- Resulting in a prototype
- Including automatic
  - cleaning of the retrieved web pages
  - filtering
  - relevance ranking
  - classification
  - keyword assignment
- Interface for browsing, searching, visualising the resulting collection
- See <http://www.jrc.it/langtech/osilia.html>
- JRC report: Scheer et al. 2000



- Extraction of a variety of information aspects from texts
- The **current** tool set can extract the following information:
  - text language (bigram-based, statistical tool; 99.7% success rate on 8000+ texts)
  - keywords (based on lemma frequency, statistical tool)
  - Eurovoc thesaurus descriptor terms (prototype)
  - references to geographical place names (prototype)
  - most similar documents to a given one (currently only monolingual or En-Es)
  - hierarchical cluster analysis of texts (bottom-up 'classification')
  - subject domains (first experiments)
- Further information extraction **planned**:
  - references to names of people, companies, organisations
  - references to products and product groups, according to the *Customs Tariff Code* ('knife', 'tyres', 'lavatory pans', 'polymerisation products', 'Pacific halibut', ...)
  - date and currency expressions
  - ...



## Sample Text




**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.



## Document Profile (Plan)



Structured Multilingual Display of Monolingual Information

**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://cnnn.com/digitaljam/wires/9906/13/plutonium\\_eu.html](http://cnnn.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**

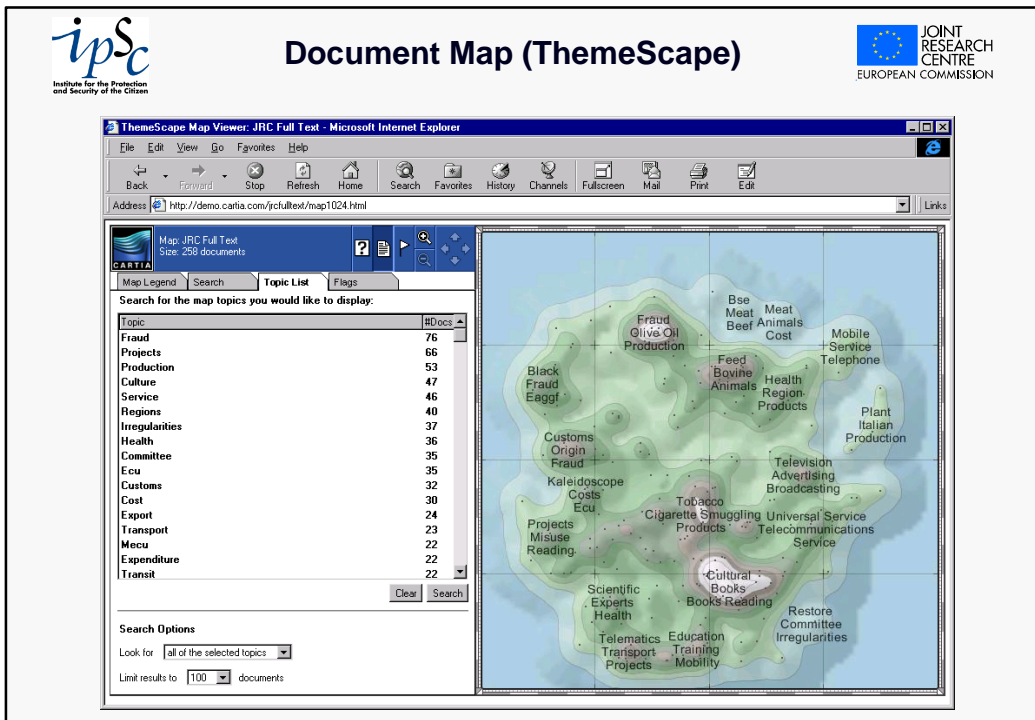
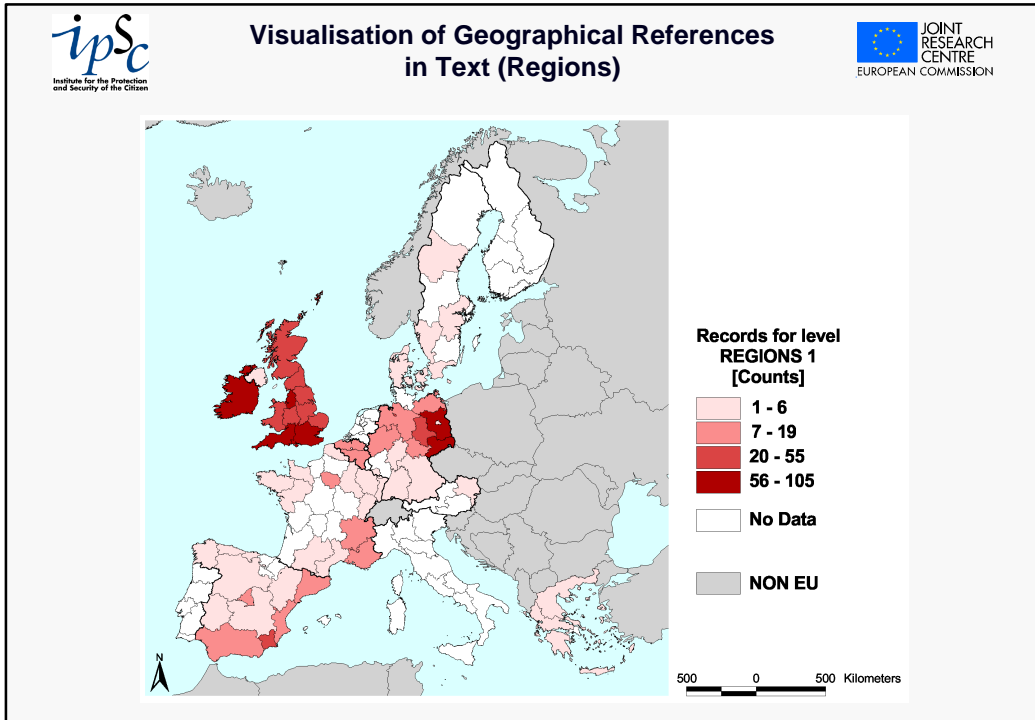
**E-3083/95 by Martin Schulz (PSE)**


**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%
<b>Others:</b>	0%






### Monolingual Keyword Identification (Indexing)




- Useful to give users an overview of the approximate contents of a document
- Necessary component for [cross-lingual Eurovoc thesaurus indexing](#)
- Statistical tool identifies *statistically salient* words of the text (same language)
- By comparing the lemma frequency of a document with an 'expected' / average lemma frequency (reference corpus frequency)
- 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
...			



### Cross-lingual Thesaurus Indexing



- IPSC Exploratory Research project 'Cross-lingual 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.

**English Text**

Resolution on radioactive waste

➔

FUEL REPROCESSING

RADIOACTIVE WASTE

NUCLEAR FUEL

PLUTONIUM

NUCLEAR SAFETY

TRANSPORT OF DANGEROUS GOODS

NUCLEAR NON-PROLIFERATION

RADIOACTIVE MATERIALS

EAEC

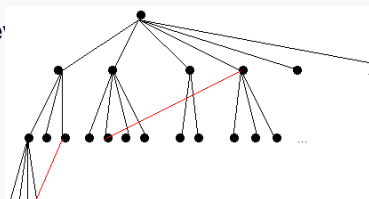
STORAGE

## 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**) !
- 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)

- 04 Politics
- 08 International Relations
- 10 European Communities
- 12 Law
- 16 Economics
- 20 Trade
- 24 Finance
- 28 Social Questions
- 32 Education and Competition
- 36 Science
- 40 Business and Competition
- 44 Employment and Working Conditions
- 48 Transport
- 52 Environment
- 56 Agriculture, Forestry and Fisheries
- 60 Agri-Foodstuffs
- 64 Production, Technology and Research
- 66 Energy
- 68 Industry
- 72 Geography
- 76 International Organisations

- 28 SOCIAL QUESTIONS**
- 2806 family
- 2811 migration
- 2816 demography and population
- 2821 social framework
- 2826 social affairs
- 2831 culture and religion**
- arts
- cultural policy
- culture**
- acculturation
- civilization
- cultural difference
- cultural identity**
- RT: protection of minorities (1236)
- RT: socio-cultural group (2821)
- cultural pluralism
- popular culture
- regional culture
- religion
- 2836 social protection
- 2841 health
- 2846 construction and town planning

## Motivation for Thesaurus Indexing

- Large organisations use thesauri to index the documents in their archives for storage and retrieval
- Advantages
  - consistency of choice of 'keywords' facilitates retrieval
  - multilinguality
- Disadvantages
  - little flexibility (descriptors like 'internet', 'Kosovo' introduced late)
  - very costly human process
  - quality and choices of human descriptor assignment are very variable

## Thesaurus Indexing vs. Using Existing Search Engines


### Thesaurus Indexing

- Can be used to retrieve documents
- **Gives users an idea of the major document contents**
- Used for given document set
- Possibility of **multilingual search + display**
- Human thesaurus indexing is **expensive** (human labour)
- Thesauri are less flexible (updating of thesaurus)
- Easier to expand search automatically to **find synonyms and hyponyms** (NTs)


### Search Engines

- Can be used to retrieve documents
- Search tool only
- Used for **open text collections** (internet)
- (Usually) monolingual search
- Cheap, once system has been developed
- More **flexible**
- User has to think of synonyms and hyponyms (NTs)

- ➔ Both tools serve a different purpose. Use both!
- ➔ Aim of automating thesaurus indexing: help or replace human thesaurus indexing




## Assigning Eurovoc Thesaurus Descriptors to Texts




- **Challenge:** Descriptor terms like DEMOGRAPHY AND POPULATION OR CONSTRUCTION AND TOWN PLANNING are unlikely to occur verbatim 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



- Use training set (texts and manually assigned thesaurus descriptors)
- Minimal linguistic text **pre-processing**
  - lemmatisation (base form reduction of words)
  - mark-up of most frequent multi-word terms
    - e.g. *power\_plant, United\_Nations, migratory\_fish\_stock, New\_York, ...*
  - 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, ...* (Kilgariff 1996)
- **Result:** a **list of keywords** (associates) + their **keyness** (weight; relevance for the contents of the meta-text)
- **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

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

**Term Frequency, Inverse Document Frequency**  
 Considers occurrence frequency of lemma (l) in meta-text ( $TF_{l,t}$ ) and number of descriptors (d) for which the lemma is an associate ( $DF_l$ )

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

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

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


**Okapi** considers occurrence frequency of lemma as an associate ( $DF_l$ ); 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)

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


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

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

'**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  
**F622** 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	F622	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

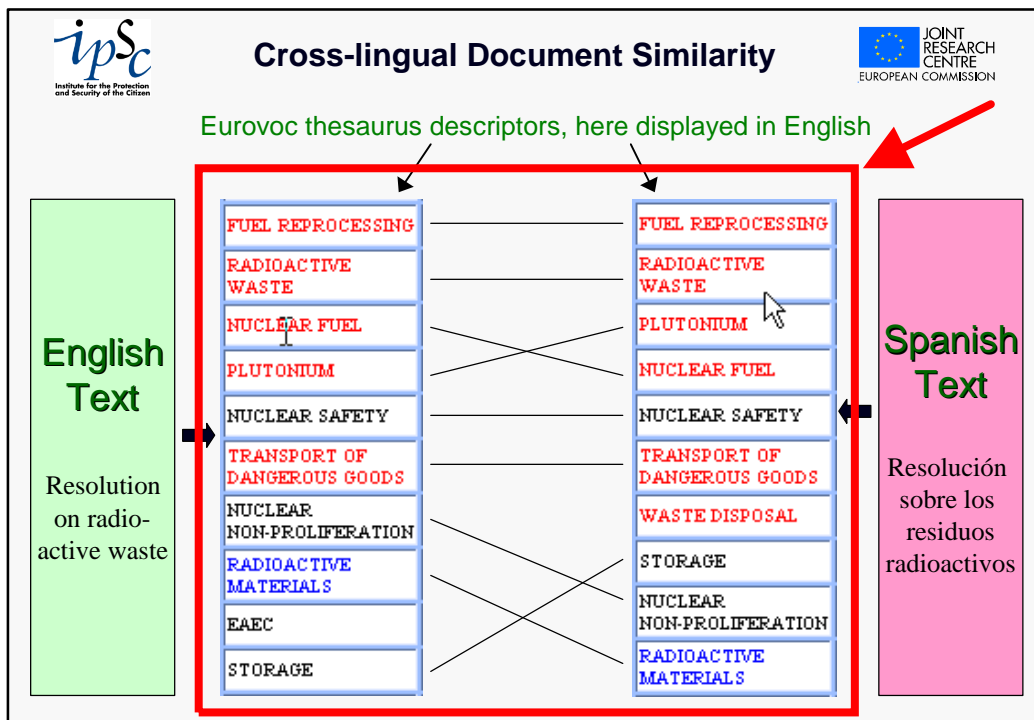
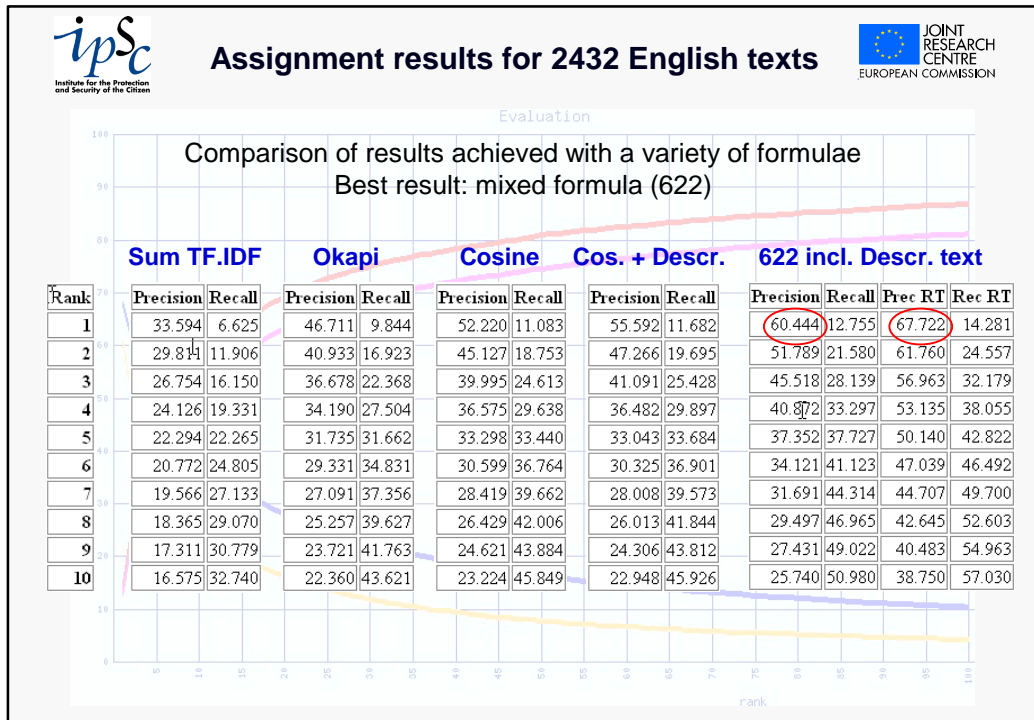



### Evaluation of the assignment



- 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.






**ipSc**  
Institute for the Protection  
and Security of the Citizen

### (Cross-language and Monolingual) Document Similarity Calculation




JOINT  
RESEARCH  
CENTRE  
EUROPEAN COMMISSION

- **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)



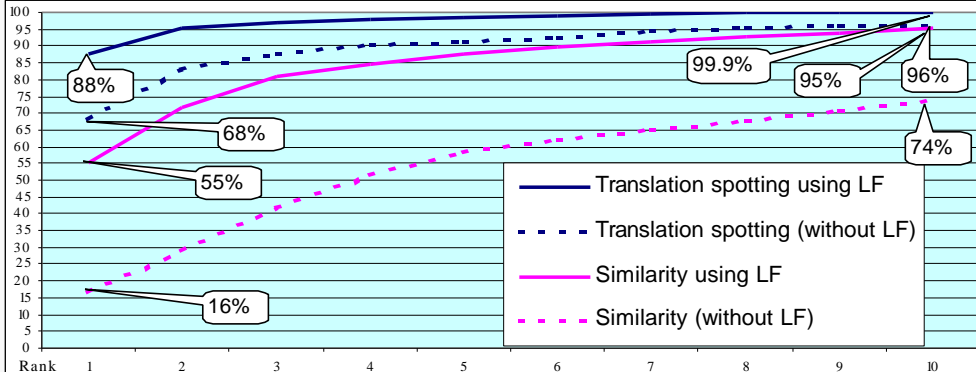
**ipSc**  
Institute for the Protection  
and Security of the Citizen

### Results for SC and Translation Spotting



JOINT  
RESEARCH  
CENTRE  
EUROPEAN COMMISSION

- **Task:** find Spanish translations of 920 English texts
- **Search space:** Collection of 2995 English and 2995 Spanish documents (total: 5990 texts)
- Average score of all translation pairs: 69%; s.d.: 0.125
- relatively low *ranking* of translations due to many similar documents (slight variations of texts) so that same-language documents seemed more similar



Rank	Translation spotting using LF (%)	Translation spotting (without LF) (%)	Similarity using LF (%)	Similarity (without LF) (%)
1	88%	55%	68%	16%
10	99.9%	95%	96%	74%



## Limitations to Statistical Thesaurus Indexing



- 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
- Usage for in-house multilingual document retrieval?
- Apply to real world scenarios
  - apply to more languages
  - incorporate with a working system at customer sites