




Automating the Assignment of Eurovoc Descriptors to text

Eurovoc Conference 2003
European Parliament
Brussels, Friday, 7 March 2003

Ralf Steinberger & Bruno Pouliquen


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


- Who we are and what we do
- Overview of our approach to Eurovoc indexing
- More detailed explanation of our approach
- Evaluation procedure and results for the automatic assignment
- Some problems we encountered, and possible solutions
- Why I am here: our offer and request
- Automatic Eurovoc descriptor assignment will not replace the work of professional indexers, but could help them.



JRC Sites in Europe



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
 - ...



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


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.



JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

Structured Multilingual Display of Monolingual Information



Joint Research Centre
European Commission


Document Profile

Display Language: English
(En, Fr, De, Es, It, Pt, Gr, El, He, Mi, Sl)

Title: Seizure of plutonium at Munich airport (E-308395)	Retrieval Date: 03.08.1994
Author: Martin Schulz (PSE)	Creation Date: 23.05.1993
Text Language(s): English	Text Length: 387 words
Source: http://mrhc.com/olymp/olympwww/eurovoc/308395plutonium_eu.html	
Related Documents: 12 (click here to view)	

<p>Keywords (Occurrence Frequency)</p> <p>TU (3), Commission (7), Karlsruhe (3), seizure (5), EU (2), plutonium (3), suitcase (3), German (4), material (4)</p>	<p>Eurovoc Thesaurus Descriptors</p> <p>plutonium, import, Hilt trade, Federal Republic of Germany, EAEC Joint Research Centre, airport, fraud</p>
--	---


<p>Names</p> <p>Organisations: Commission, European Institute for Transuranium Elements (TUE), Joint Research Centre, PSE</p> <p>People: Martin Schulz, Mrs. Breyer, Mr. Papadakis</p> <p>Geographical References:</p> <p>Germany (1) click here to view</p> <p>Germany (1), Karlsruhe (3), Munich (2), Germany (1), Federal Republic of Germany (1)</p> <p>No Others</p>	<p>Combined Nomenclature Product Groups</p> <p>CN 2844: "radioactive chemical elements and radioactive isotopes, not in bulk or form chemical elements and isotopes, and their compounds, mixtures and residues containing these products" (plutonium; 3)</p> <p>CN 4204: "Trunks, suit, vanity, executive, brief, spectacle, binocular, camera, music instrument, gun cases, trunks and similar, traveling, toiletry bags, rucksacks, handbags, school satchels, shopping-bags, wallets, purses, map, cigarette cases" (suitcase; 3)</p>	<p>Document Summary</p> <p>E-308395 by Martin Schulz (PSE)</p> <p>Seizure of plutonium at Munich airport</p> <p>In the summer of 1984 a suitcase containing plutonium (legally imported into Germany) was seized in a terminal at Munich airport in the Federal Republic of Germany. The Commission (European Safeguards Directorate) was alerted by the German authorities in the early afternoon of 10 August, 1994, that some material might be seized.</p> <p style="text-align: right;">See full text</p> <p style="text-align: right; font-size: small;">http://www.ec.europa.eu/olymp/</p>
---	--	---




JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

Agenda

- Who we are and what we do
- Overview of our approach to Eurovoc indexing
- More detailed explanation of our approach
- Evaluation procedure and results for the automatic assignment
- Some problems we encountered, and possible solutions
- Why I am here: our offer and request




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**)







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 .





JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

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 .






JOINT RESEARCH CENTRE
EUROPEAN COMMISSION


Library Usage vs. JRC Usage

- Automatic assignment is an approximation, a "best guess".
- Achievable quality is clearly lower than that of human assignment. Libraries need high quality for indexing and retrieval.
- Requirements differ.
- JRC also wants to
 - index, retrieve, and give cross-lingual information access.

+

- We index documents that would otherwise not be indexed at all.
- Cross-lingual **document similarity** calculation
- Multilingual **classification**
- Multilingual **clustering** of documents
- Multilingual **document maps**
- Subject-specific **summarisation**








Identify Statistically Salient Words in Text

- Compare the word frequency (**lemma**) in a document (**TF**) with an 'expected' / average word frequency (reference corpus frequency **RCF**)
- using the statistical *log-likelihood* test (Dunning 1993), or others.
- **Text length:** 300 words; **Reference corpus length:** 100 million words

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 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

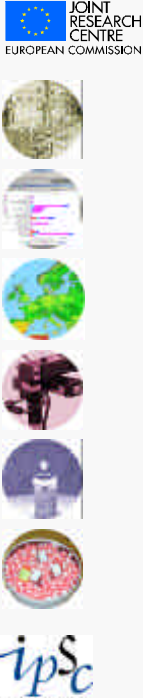
→ **Result of Training:** Weighed associate lists for all descriptors



Associate List: RADIOACTIVE MATERIALS

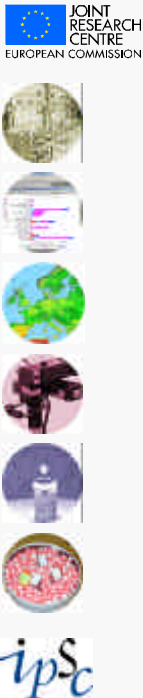
deuterium	35.7836791092845
lithium	33.0805724769899
thorium	32.560703225522
tritium	32.0826451843048
nuclear_material	13.79399100837
radioactive_material	7.84970673161556
plutonium	6.72955494180221
radioactive_substance	6.43422856440347
nuclear	5.851612117697
undine_uta_bloch_von_blottnitz	5.53278869694883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

Associate List: FISHERY MANAGEMENT




fishery-related	fishery_resource	54.4721542368865
	fishing	49.111563204862
	fish	46.1954956023147
	common_fishery_policy	44.6741843971235
	fishary	44.1911518447189
	fishing_activity	43.9277571334009
	fly_the_flag	42.8744724342378
	aquaculture	39.2740718215554
	conservation	38.3480454820621
	vessel	37.911138722495
management-related	fishing_vessel	37.8343365844953
	catch	36.8503034704154
	fish_stock	34.9289335973103
	face	34.388433383343
	allowable_catch	33.2880590561664
	catch_quote	32.2683540654092
	control_system	31.1753892078216
	fish_for	29.836606340017
	nautical_mile	29.541951528159
	fishing_right	29.1916760888821
centimetre	28.7167311460535	
control_measure	28.0527345432075	
gross_tonnage	26.0043818725124	
fishing_zone	27.8678836557102	

Associate List: MAURITANIA




fishery-related	mauritania	26.5901338018222
	islamic_republic_of_mauritania	21.5028701117865
	pole_and_line_tuna	7.5756159727615
	mauritanian	7.38435884830453
	fishery_sector	6.82371824209186
	cephalopod	6.23513420563961
	datasheet	6.12232127368836
	shipowner	5.50033686503602
	coast	5.13094774515327
	vae	4.50387082013722
sea-related	dispose_to	4.15622870775623
	onnum	4.09479477600525
	inoulchott	3.97236867452946
	fishing	3.86934015616878
	fishing_zone	3.93325407502887
	much_oblige	3.81347622659801
	surface_longliner	3.89812021954454
	pelagic	3.82550430535037
	observer	3.60326786424058
	application_form	3.50839353911244
fishery_agreement	3.49778120811487	
tomngil	3.47804638502223	
fec	3.34136820199739	
fish	3.31932642852838	



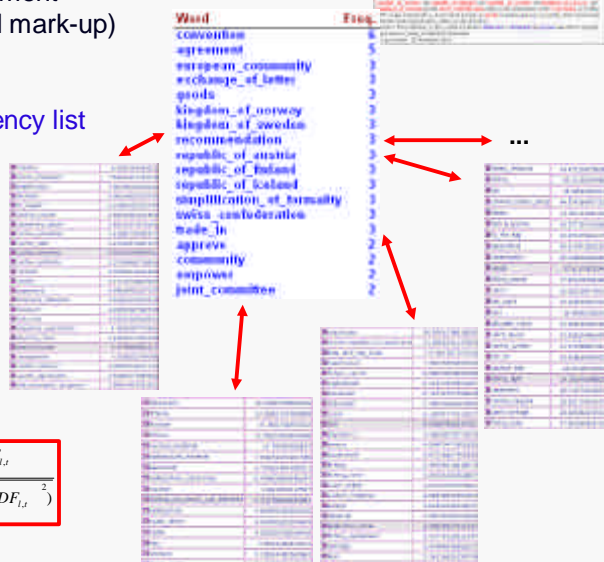
JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

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

Word	Freq
convention	5
agreement	5
european_community	3
exchange_of_letters	3
goods	3
kingdom_of_norway	3
kingdom_of_sweden	3
recommendation	3
republic_of_austria	3
republic_of_finland	3
republic_of_italy	3
simplification_of_formality	3
swiss_confederates	3
made_in	2
approve	2
community	2
empower	2
joint_committee	2



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







JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

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 ²	Cosine	Rank Cosine	Okapi	Rank Okapi	Rank	Prec	Rec
5641040706000000	FISHING CONTROLS [g]	.000144033	0.360	1	95.169	1	3	100	30
5641020000000000	FISHING OPINIONS [H]	.00243464	0.308	2	65.038	14	2	100	20
5641040200000000	COMMON FISHERIES POLICY [g]	.00038023	0.280	3	52.910	20	3	100	30
5641040300000000	FISHERY MANAGEMENT [H]	.000207086	0.270	4	79.262	6	4	100	40
5641040700000000	FISHING REGULATIONS [g]	.000107034	0.270	5	79.982	5	5	100	50
5641040704000000	FISHING PERMIT [g]	.00306631	0.261	6	71.577	8	6	100	60
5641040301000000	CONSERVATION OF FISH STOCKS [H]	.000180818	0.252	7	83.982	3	7	85	60
5641040600000000	FISHING AREA [g]	.000182474	0.252	8	84.178	2	8	75	60
5601604030000000	CONSERVATION OF RESOURCES [H]	.000234809	0.231	9	55.311	26	9	66	60
5641050000000000	FISHERY RESOURCES	.000400863	0.232	10	75.046	7	10	60	60
5641040600000000	CATCH OF FISH	.000313101	0.215	11	67.687	9	11	54	60
5641040000000000	FISHERIES POLICY	.00258399	0.205	12	58.416	23	12	50	60
5641040705000000	FISHING LICENCE	.000371136	0.181	13	57.818	25	13	46	60
5641060300000000	FISHING FLEET	.00100478	0.179	14	63.323	19	14	42	60
5641010000000000	FISHING INDUSTRY	.000551953	0.176	15	39.289	42	15	40	60
5641040201000000	EURORICHEL	.000706822	0.176	16	62.240	21	16	37	60






JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

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 (idf ²)	Cosine =	Rank Cosine	Okapi	Rank Okapi	Rank	Prec	Rec
ES41040706000000	FISHING CONTROLS [R]	.001144099	0.360	1	35.169	1	1	100	10
ES41040708000000	FISHING GROUND [H]	.00243464	0.306	2	65.016	14	2	100	20
ES41040800000000	COMMON FISHERIES POLICY [G]	.00038823	0.280	3	62.930	20	3	100	30
ES41040100000000	FISHERY MANAGEMENT [M]	.000287080	0.279	4	79.362	6	4	100	40
ES41040700000000	FISHING REGULATIONS [S]	.000197034	0.270	5	79.982	5	5	100	50
ES41040704000000	FISHING PERMIT [C]	.00306501	0.261	6	71.677	9	6	100	60
ES41040301000000	CONSERVATION OF FISH STOCKS [H]	.000189818	0.255	7	83.982	3	7	85	60
ES41040800000000	FISHING AREA [O]	.000182474	0.252	8	84.178	2	8	75	60
ES20604030000000	CONSERVATION OF RESOURCES [M]	.000234209	0.251	9	55.311	26	9	66	60
ES41050000000000	FISHERY RESOURCES	.000403863	0.232	10	75.046	7	10	60	60
ES41040800000000	CATCH OF FISH	.000513101	0.215	11	67.667	9	11	54	60




JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

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

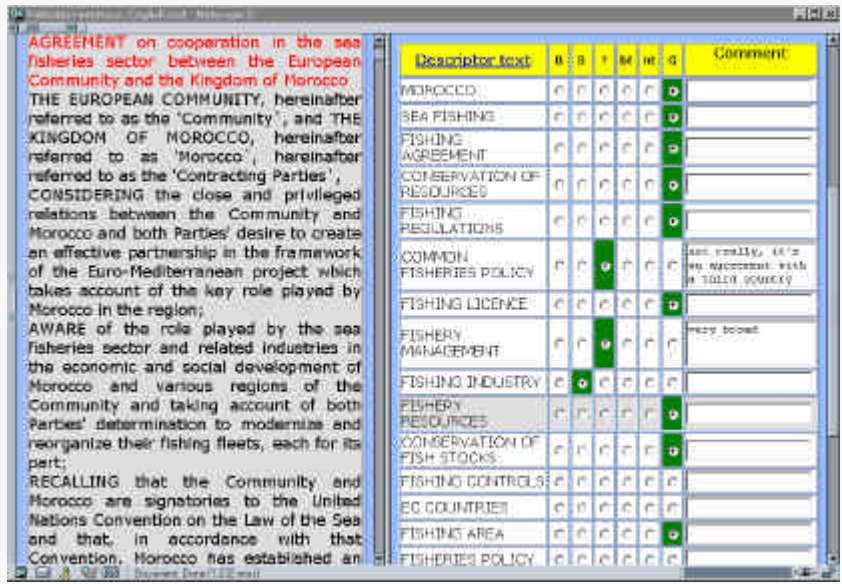


JOINT RESEARCH CENTRE
EUROPEAN COMMISSION


Manual Evaluation of the Assignment



ipsc




The screenshot shows a manual evaluation interface. On the left, there is a text document titled 'AGREEMENT on cooperation in the sea fisheries sector between the European Community and the Kingdom of Morocco'. The text describes the agreement between the European Community and Morocco regarding fisheries cooperation. On the right, there is a table with columns for 'Descriptor text', 'B', 'S', 'Y', 'M', 'H', 'G', and 'Comment'. The table lists various descriptors such as 'MOROCCO', 'SEA FISHING', 'FISHING AGREEMENT', 'CONSERVATION OF RESOURCES', 'FISHING REGULATIONS', 'COMMON FISHERIES POLICY', 'FISHING LICENCE', 'FISHERY MANAGEMENT', 'FISHING INDUSTRY', 'FISHERY RESOURCES', 'CONSERVATION OF FISH STOCKS', 'FISHING CONTROLS', 'EC COUNTRIES', 'FISHING AREA', and 'FISHERIES POLICY'. Each descriptor has a set of checkboxes corresponding to the columns B, S, Y, M, H, G, and a 'Comment' column.

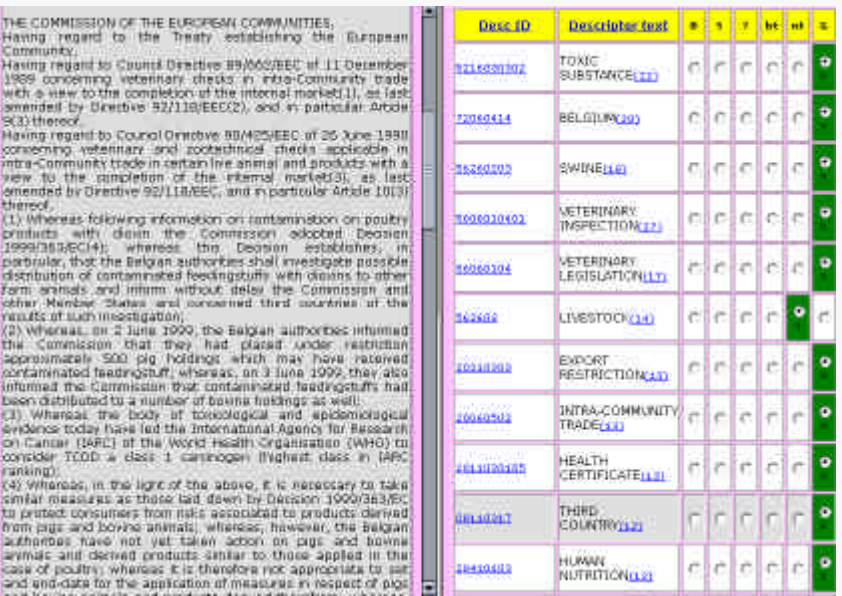


JOINT RESEARCH CENTRE
EUROPEAN COMMISSION


Eurovoc Assignment – Discussion (1) (Good Result)



ipsc



The screenshot shows a Eurovoc assignment interface. On the left, there is a text document titled 'THE COMMISSION OF THE EUROPEAN COMMUNITIES, Having regard to the Treaty establishing the European Community...'. The text discusses veterinary checks and trade in intra-Community trade. On the right, there is a table with columns for 'Desc. ID', 'Descriptor text', 'B', 'S', 'Y', 'M', 'H', 'G', and 'Comment'. The table lists various descriptors such as 'TOXIC SUBSTANCE(10)', 'BELGIUM(20)', 'SWINE(16)', 'VETERINARY INSPECTION(17)', 'VETERINARY LEGISLATION(17)', 'LIVESTOCK(14)', 'EXPORT RESTRICTION(43)', 'INTRA-COMMUNITY TRADE(10)', 'HEALTH CERTIFICATE(11)', 'THIRD-COUNTRY(12)', and 'HUMAN NUTRITION(13)'. Each descriptor has a set of checkboxes corresponding to the columns B, S, Y, M, H, G, and a 'Comment' column.



JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

Eurovoc Assignment – Discussion (2) (Bad Result)

Whereas the Commission considers that the assessors' study provides sufficient assurances that the estimated growth figures are reliable,
HAS ADOPTED THIS DECISION:


Article 1:
The seat capacity restrictions imposed on Aer Lingus by Decision 94/118/EC of 21 December 1993 in State aid No C 34/93, for the routes between Ireland and the United Kingdom and Dublin and London-Heathrow, are adjusted according to Article 1 (f) of the Decision to reflect the corresponding market growth in 1994 and 1995, as follows:

- DUB-LHR:
- 1994: 1 451 821 seats,
- 1995: 1 621 684 seats,
and
- IRL-LUK:
- 1994: 3 570 765 seats,
- 1995: 3 845 714 seats.

Article 2:
This Decision is addressed to Ireland.
Done at Brussels, 30 November 1994.
For the Commission,
Manuelino ORFILA
Member of the Commission

(1) OJ No L 54, 25. 2. 1994, p. 30.
</FIELDITE>


Desc ID	Descriptor text	B	S	Y	BT	NT	G
48800205	CIVIL AVIATION(11)	✓	✓	✓	✓	✓	○
10060307	EC ECONOMIC AND SOCIAL COMMITTEE(11)	○	○	○	○	○	○
10060703	EC COMMITTEE(10)	○	○	○	○	○	○
10060304	COMMITTEE OF THE REGIONS(10)	○	○	○	○	○	○
44210109	APPOINTMENT OF STAFF(10)	○	○	○	○	○	○
72060401	FRANCE(7)	✓	✓	✓	✓	✓	○
482602	AIR TRANSPORT(5)	✓	✓	✓	✓	✓	○
72060404	REPUBLIC OF IRELAND(6)	✓	✓	✓	✓	✓	○
483104	RESTRICTION ON COMPETITION(5)	✓	○	○	○	○	○
48110311	TRANSPORT CAPACITY(5)	✓	✓	✓	✓	✓	○




JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

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






JOINT RESEARCH CENTRE
EUROPEAN COMMISSION


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?





ipsc
Institute for the Production of Synthetic Chemicals



JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

Some Difficult Cases (1)

- Some descriptors are irrelevant, although many good associates were found, e.g. plutonium sample text

Automatically assigned descriptors:

Descriptor text	Quality
FEDERAL REPUBLIC OF GERMANY	0.312
PLUTONIUM	0.304
PLUTONIUM SAMPLE	0.304
PLUTONIUM SAMPLES	0.290
PLUTONIUM SAMPLES	0.281
PLUTONIUM	0.192
PLUTONIUM	0.188
PLUTONIUM MATERIAL	0.173
COMPOSITION OF PLUTONIUM	0.156
PLUTONIUM	0.154

Individual Trigger Associates:

Learning: *Plutonium in text*

- plutonium
- plutonium

Learning: *Plutonium in text*

- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium

Learning: *Plutonium in text*


- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium


Learning: *Plutonium in text*

- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium
- plutonium

Manually assigned descriptors:

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





ipsc
Institute for the Production of Synthetic Chemicals

Some Difficult Cases (2)




- Wrong assignment of semantically related descriptors with similar associate lists:
RADIOACTIVE WASTE vs. **TRANSPORT OF DANGEROUS GOODS**

Lemma	Lemma
radioactive_waste	dangerous_goods
nuclear_waste	radioactive_material
seffafield	by_road
nuclear_fuel	carriage
nuclear	dangerous
undine_uta_bloch_von_blottnitz	plutonium
radioactive	radioactive_waste
nuola_aheri	nuclear_fuel
reprocess	shipment
dounney	ack
nuclear_material	bind_for
waste	tank
shipment	receptacle
fuel_element	transport
nuclear_power_station	pollute
radioactive_material	nuclear_waste

Current State of our Work (1)

- System is currently optimised for **English and Spanish**
- System is trained for another eight languages without pre-processing:
De, It, Pt, NI, Da, Sv, Fi; Fr with using stop words only

Language	Without linguistic pre-processing	With pre-processing
En	~55	~62
Es	~48	~58
Da	~52	~52
De	~50	~50
Fi*	~50	~50
Fr	~52	~60
It	~53	~53
NI	~51	~51
Pt	~52	~52
Sv*	~50	~50






Current State of our Work (2)

- We work with Eurovoc version 3.1

- In English, associate lists exist for **2565 descriptors**

- Many Eurovoc descriptors are not used.
In the over 25000 training texts of half a page or more:
 - **35% have never been used!**
 - 9% were only used once
 - 23% were only used 2-8 times

- ➔ We cannot assign descriptors without training material.



Our Offer and Request

- We would like our software tool to be actively used.
 - Fully automatic assignment for texts that would not be indexed at all.
 - As an interactive tool to support human indexers.

- We need feedback on our work.
- We need more **manual evaluation** of automatic assignment results.
 - Any volunteers? Please contact me at:

Ralf.Steinberger@jrc.it
Tel: +39 – 0332 786271
<http://www.jrc.it/langtech>

