

Using Language Technology to support the fight against fraud

23 May 2000

Ralf Steinberger

EC - Joint Research Centre

Institute for Systems, Informatics and Safety (ISIS)
Risk Management and Decision Support Unit (RMDS)
Anti-fraud Information Management Sector (AIM)

- What is Language Technology (LT)?
- Difficulties in dealing with language automatically
- How can LT support the fight against fraud?
- Goals of the LT activities of the
Anti-fraud Information Management sector
- Extraction of information from texts: keywords, geographical references, document and word similarity, ...
- Visualisation of the extraction results: document profile, document map, map of geographical references mentioned in texts, ...

- **Computational Linguistics (CL)** is a field in the cross section between computers and natural language. It draws its main knowledge from linguistics, computer science and statistics.
- **Natural Language Processing (NLP), Language Engineering (LE), Language Technology (LT)** are similar terms referring to the application areas of CL.
- Application areas: **Machine Translation (MT)**, Spell checking and grammar checking, Speech Recognition, **Information Retrieval (IR)**, Summarisation, **Document Classification**, (Personalised) **Information Filtering (PIF)**, Keyword Extraction, Language Recognition, Computer-Assisted Language Learning (CALL), ...

- Word forms: *aller, va, iriez, aille, allait, allâtes, allant, allées, ...*
- Word sense ambiguity: *bar* (*metal rod, stripe, drinking place, bar of lawyers, ...*) → translational ambiguity
- Multi-word expressions: *colour liquid crystal display* vs. *Flüssigkristallfarbbildschirm*
- Syntactic ambiguity:
 $[Time]_{SU} \text{ flies}_V [like\ an\ arrow]_{ADV}$
 $[Time\ flies]_{SU} \text{ like}_V [an\ arrow]_{OBJ}$
- Only some information is contained in language:

The man saw the elephant with the telescope. (tool to see)

The elephant saw the man with the telescope. (attribute of man)

Language Technology can help anti-fraud agencies

- to get quick access to information ‘hidden’ in large amounts of texts, written in a variety of languages.
- to keep abreast of developments
by monitoring the web or intranets continuously,
by pointing out that some new relevant information is available
by producing summarising reports automatically.
- Important for EC: multilinguality!

- **Cross-language Document Retrieval**
 - agent which retrieves documents in a variety of languages which match query words (✓, –)
- **Information Extraction (document profile)**
 - language recognition, free indexing terms, geographical references, document similarity (✓)
 - Eurovoc multilingual indexing terms, names, products, summary, subject domains, ... (–)
- **Information Visualisation**
 - document profile (✓, –)
 - document maps (✓)
 - geographical representation (✓)
 - representation of extracted information in tables (–)
 - show trends (–)

Text

See page 1 of
Audit_ISIS_LE_990610_part2.doc

See page 2 of
Audit_ISIS_LE_990610_part2.doc

Bigram Analysis of Ten-Langs.txt - Microsoft Internet Explorer

File Edit View Go Favorites Help

[1: Komissio esittää yhteenvedon petostentorjuntaa koskevista toimistaan yhteisön taloudellisten etujen suojaamista käsittelevässä vuosikertomuksessa. Eurooppalaiset veronmaksajat vaativat julkisten varojen moitteetonta käyttöä ja petoksilta suojaavaa politiikkaa]

[2: Årsrapporten om skydd av gemenskapens finansiella intressen visar vad kommissionen har gjort under året i kampen mot bedrägerier och fusk. Skattebetalarna i EU kräver att deras pengar skall användas till rätt ändamål och de förväntar sig att verksamheten skall gå säker från bedrägerier]

[3: I 1995 blev der gjort væsentlige fremskridt på lovgivningsområdet med vedtagelsen af to grundlæggende tekster, nemlig en om kontrol og administrative sanktioner (horisontal forordning), og en om en strafferetlig beskyttelse af fællesskabets finanser (konvention)]

[4: Auf operationeller Ebene zeigen die auf Initiative der Kommission in Zusammenarbeit mit den nationalen Behörden durchgeführten Ermittlungen immer deutlicher, daß die Kommission auch in der Arbeit vor Ort ihren Zusatznutzen bei der Betrugsbekämpfung einbringen kann]

[5: Daarom is zij van mening dat deze mogelijkheden bij gelegenheid van de Intergouvernementele Conferentie (ICC) moeten worden uitgebreid. Zij heeft dan ook de nodige initiatieven genomen om deze kwestie op de agenda van de ICC, die in maart 1996 begint, te doen plaatsen]

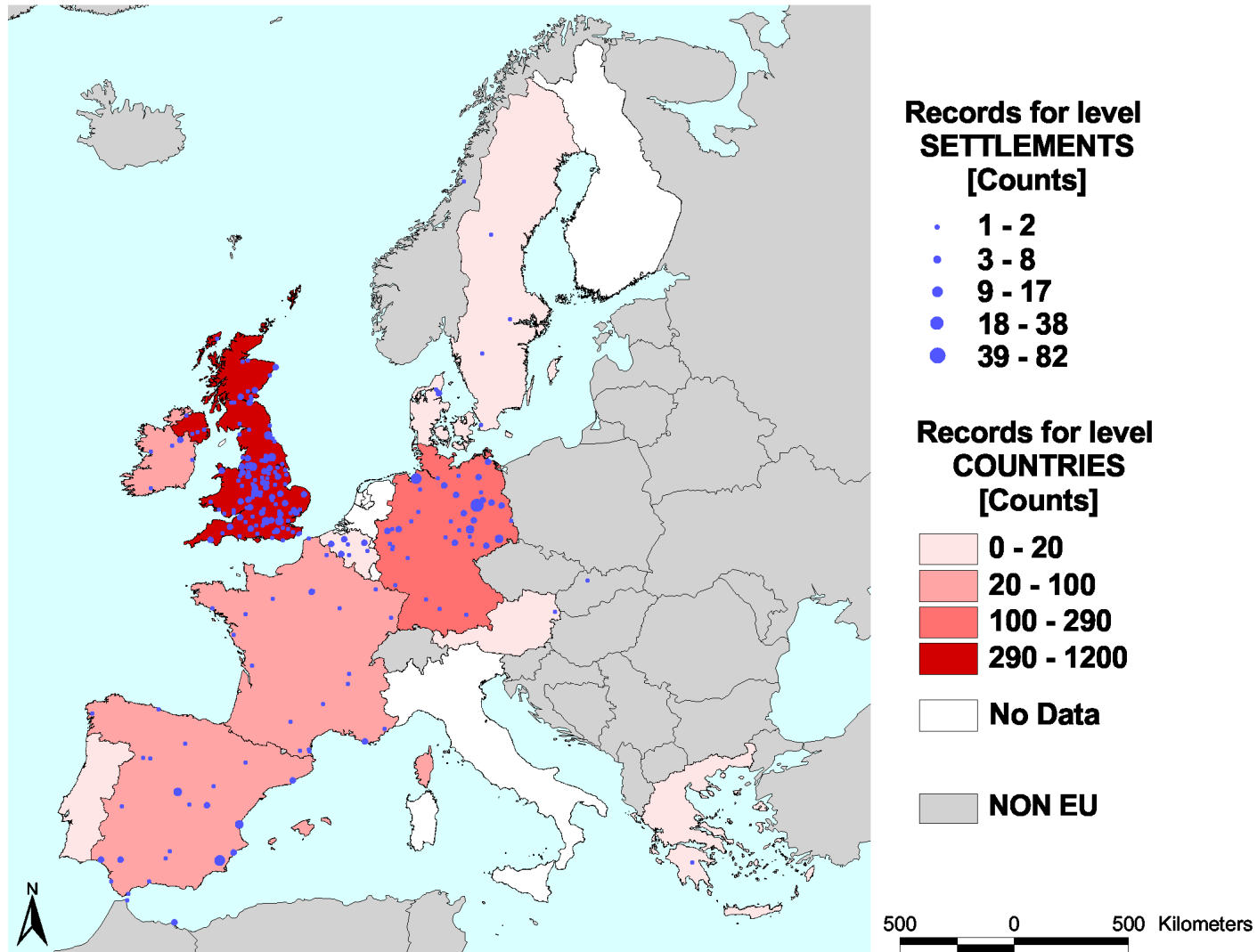
[6: The scale of the amounts at stake illustrates how the launching of in-depth inquiries in cooperation with specialized departments in the Member States as soon as fraud is suspected can reveal the existence of networks, often criminal, operating with highly sophisticated techniques]

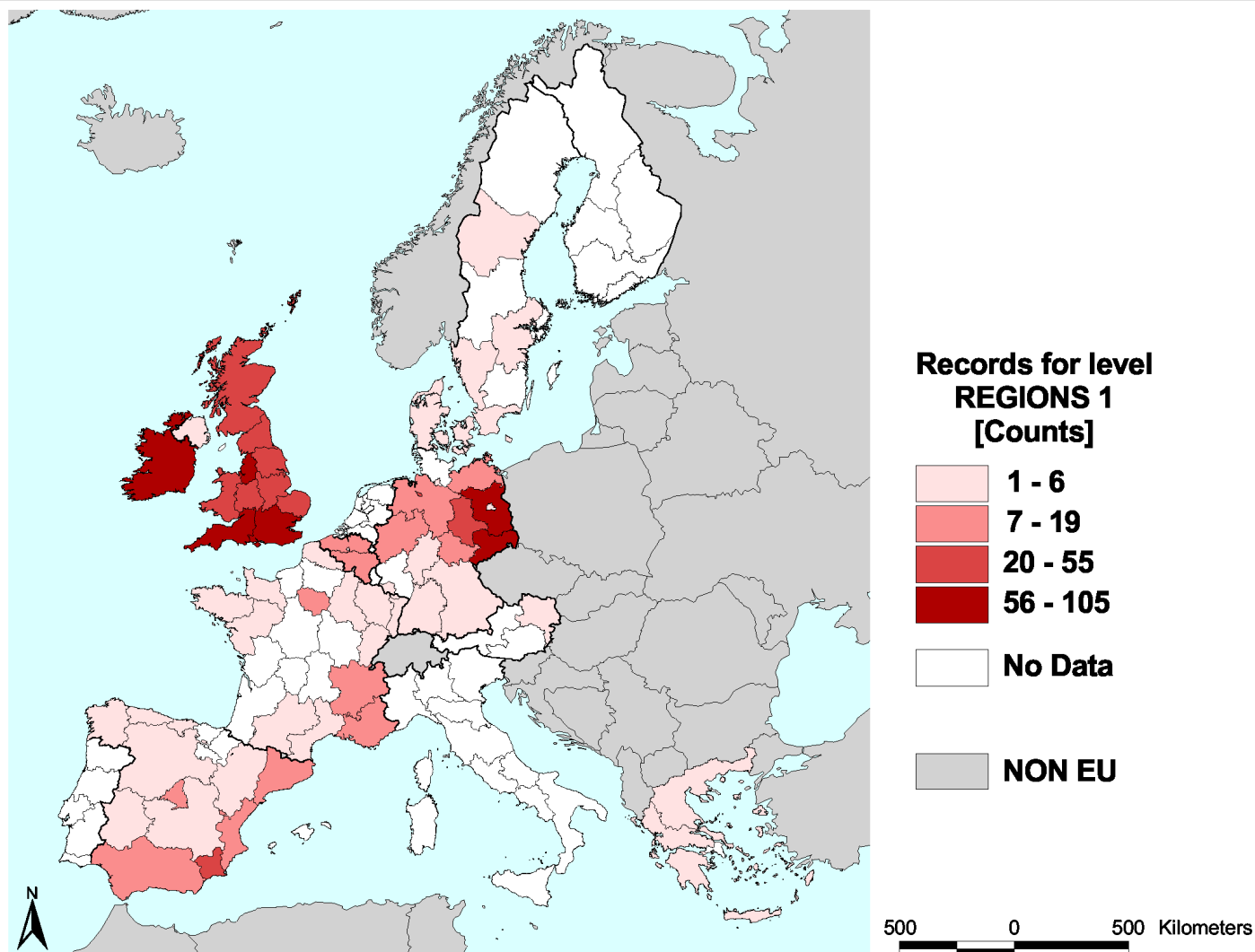
[7: La proposition initiale de ce règlement relatif à la protection des intérêts financiers, présentée par la Commission en juillet 1994, prévoyait, au-delà des mécanismes de sanction administrative communautaire, des règles générales pour les mesures et les contrôles communautaires]

[8: Benché sostanziali progressi siano stati compiuti sul piano legislativo, l'esperienza ha dimostrato, a giudizio della Commissione, che gli strumenti giuridici offerti dal Trattato sono insufficienti tenuto conto della tutela che esigono le vaste risorse del bilancio dell'Unione]

[9: Esta concepción horizontal señala una orientación clara y decisiva en el enfoque legislativo de la lucha contra el fraude, y representa asimismo un paso importante hacia una mayor coherencia del marco legislativo en materia de protección de los intereses financieros comunitarios]

[10: Por último, no que diz respeito ao regime de co-financiamento comunitário destinado ao reforço dos controlos dos Estados-membros, a Comissão elaborou um relatório sobre a situação da aplicação do actual regime com vista a formular novas propostas destinadas a colmatar as lacunas do antigo regulamento]





<u>KEYWORD</u>	<u>KEYNESS</u>
----------------	----------------

TUI	65.31
Commission	62.27
Karlsruhe	57.55
seizure	55.84
OJ	42.21
plutonium	39.78
suitcase	38.44
German	29.49
material	28.51
Munich	23.60
Breyer	22.52
airport	17.80

<u>KEYWORD</u>	<u>KEYNESS</u>
----------------	----------------

PSE	17.06
Schulz	16.46
Euratom	15.99
Joint	14.11
Germany	12.83
authority	11.79
directorate	11.78
answer	11.58
question	11.56
safeguards	11.05
sensational	11.04
alert	10.98

- Problem of the presented method:
 - no compounds (e.g. *'power plant'*)
 - monolinguality
 - inconsistency ('bread' vs. 'toast' vs. 'bakery products'),
- Human indexers often use 'controlled vocabulary' such as Eurovoc (The EP's Eurovoc thesaurus exists in 11 languages)
- We intend to assign such keywords automatically by training our system on manually indexed documents.
First experiments have yielded positive results.
- Advantage: consistency + multilingual indexing

Link documents written in different languages by:

- representing document contents in a language-neutral way
- i.e. linking them to the same multilingual classifications, e.g.
 - *Eurovoc* thesaurus
 - *TARIC* product nomenclature (Customs Tariff Code)
 - *Lenoch* subject codes
 - ...

document name	node+attraction	node#	docs	word#1	word#2	word#3	...
agricultural_policy_h...\ 53.\		7	1	consumer	restoration	encephalopathy	...
consumer_movement_h...../ 43.\		333	2	consumer	labelling	spongiform	...
investment_aid_h...../ 29-\		69	1	consumer	labelling	transparency	...
community_control_h...../ 22----\ 62.\		379	3	consumer	spongiform	encephalopathy	...
goat_h.....\ 62.\		166	1	processing	encephalopathy	spongiform	...
press_h...../ 42...../ 74		449	4	consumer	encephalopathy*	bovine*	...
cosmetic_product_h...../ 74		43	1	monitoring	ban	bovine	...
		471	7	bovine*	bse*	consumer*	...
		143	1	scrapie	infect	scientific	...
		304	2	scientific	scrapie	veterinary	...
		196	1	scientific	bovine	veterinary	...
		387	3	scrapie	scientific	infect	...
		74	1	scrapie	encephalopathy	infect	...

Small sample cluster of seven documents and the first three of a ranked list of indexing words for each document.

The system also calculates the most representative indexing words for each document cluster.

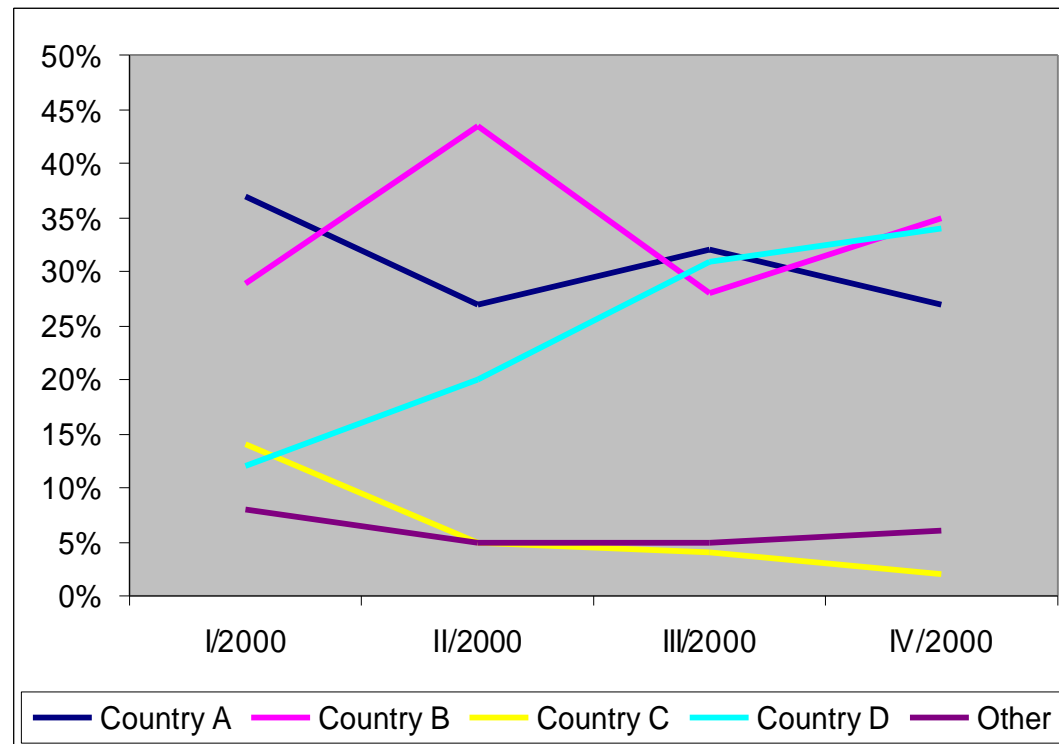
See page 3 of
Audit_ISIS_LE_990610_part2.doc

See page 4 of
Audit_ISIS_LE_990610_part2.doc

NewsMaps

as an example for a more attractive presentation to the user

Fixed parameters: product type: **tobacco**; indexing term: **fraud**
→ show the **trend for countries** mentioned in these documents
over time



Our goal:

- multilingual retrieval of potentially relevant documents from the internet, the intranet and extranets
- extraction of different information aspects from these documents; language-neutral representation where possible
- visualisation of the contents of individual documents and of document collections

Purpose:

- give cross-language access to information 'hidden' in large amounts of multilingual text
- provide automatic monitoring facility

What we have achieved so far:

- development and purchase of several standalone tools
- applied to a small number of languages
- analysis of fraud-related data for OLAF

What remains to do:

- develop and purchase further tools
- apply them to many more languages
- integrate them with each other
- prepare applications for customers
- continue thinking about how LT can serve our partners
- plan and execute projects with our partners in anti-fraud to serve their real-world needs

Pause for laughter