

Using Thesauri for Automatic Indexing and for the Visualisation of Multilingual Document Collections

Ralf Steinberger, Johan Hagman, Stefan Scheer

European Commission
Joint Research Centre
Institute for Systems, Informatics and Safety
Anti-fraud Information Management Sector
T.P. 361, I-21020 Ispra (VA), Italy
{ralf.steinberger, johan.hagman, stefan.scheer}@jrc.it
<http://www.jrc.org/>

Abstract. This article presents an approach for cross-language document comparison and for the visualisation of multilingual document collections. Document comparison usually relies on the calculation of the degree of lexical overlap between documents. As this is not possible for documents written in different languages, the contents of these documents first have to be mapped onto a language-independent representation. The JRC's statistical tool for controlled vocabulary keyword assignment assigns descriptors of the multilingual *Eurovoc* thesaurus, which can be used for cross-language document comparison. The language-independent sets of thesaurus descriptors allow to identify, for a given document, the most similar documents even if they are written in different languages. They furthermore allow to organise and to visualise the structure and approximate contents of whole multilingual document collections in two-dimensional document maps.

1 Introduction and Motivation

The European Commission's (EC) *Joint Research Centre* (JRC) is working on putting together a multilingual system which can be used by a variety of user groups to gather, analyse and view documents which match their interest profile. The system consists of three components: the first one is a crawler which automatically retrieves potentially relevant documents from the Internet and other sources; the second extracts different information aspects from these documents, including subject domains, keywords, named entities and geographical references; the third component visualises the information in a variety of ways. For both political and practical reasons, the software developed for the European Commission must be multilingual and it must allow cross-lingual access to textual data.

1.1 State of the Art in Cross-Language Software

While machine translation and cross-language document retrieval software is nowadays available commercially, there seem to be no systems which can establish automatically whether two documents written in different languages cover similar subjects or events ([1], [2]). Classification and clustering algorithms which compute the similarity between documents usually rely on their lexical overlap, but this is obviously not a solution for documents written in different languages. In order to compare documents in a multilingual text collection, the text contents need to be ‘normalised’ or ‘standardised’ somehow in a language-neutral way, for instance by mapping the document contents to a language-independent knowledge structure. Multilingual thesauri and nomenclatures are an appropriate resource to establish links between texts written in different languages. The JRC has access to several thesauri and nomenclatures existing in all eleven official European Union languages. These multilingual knowledge structures are a very valuable resource for a variety of cross-language applications. The JRC uses the *Eurovoc* thesaurus ([3]) for the assignment of keywords and for the visualisation of multilingual text collections.

1.2 The *Eurovoc* Thesaurus

Eurovoc is a thesaurus with 5.933 *descriptors* (keywords), ordered in a hierarchical structure into 21 different fields and, at the next level, into 127 *microthesauri*. There are 5877 reciprocal relations linking *broader terms* (BT) and *narrower terms* (NT) with each other. 2.730 reciprocal associations mark *related terms* (RT), and there is a language-dependent number of descriptor synonyms which are related to the descriptor by the relation *use for* (UF).

Eurovoc was developed for use by the archivists of the European Parliament (EP), the European Commission’s Publications Office (OPOCE) and other organisations as a controlled vocabulary to index¹ all documents in the archives manually. Two features make *Eurovoc* particularly suitable for the JRC’s work: (a) *Eurovoc* exists in exact, one-to-one translations in all eleven official European Union languages, and (b) the EP has given the JRC access to large amounts of manually indexed documents which can be used for the training of an automatic *Eurovoc* descriptor assignment tool. The multilingual nature of the thesaurus allows users which are speakers of one language to search a database of indexed texts for documents in another language and to get an idea of the contents of texts which are written in languages the users do not understand.

¹ In this paper, we consistently use the verb ‘to index’ in the sense ‘to assign keywords’ (i.e. a small number of words or multi-word terms which represent the contents of a document) and never in the sense *full-text indexing* (i.e. to produce an inverted index of all words occurring in a document).

2 Linking Texts to the Thesaurus

The JRC imitates the process of human indexing using an automatic procedure which is based on a learning phase in which the manually indexed documents are used, and on a statistical assignment procedure which uses the data created during the training phase.

Both training material and the new texts which are to be indexed automatically are lemmatised using Lernout & Hauspie's lemmatisation software IntelliScope® Search Enhancer² and are lower-cased before any further processing of the textual data in order to optimise the statistical procedures.

2.1 Training Phase

In the training phase, the JRC uses a monolingual indexing tool which identifies the statistically most typical ordinary language words which are related to a Eurovoc descriptor. Their cumulated occurrence in a text indicate a certain likelihood that the Eurovoc descriptor is an appropriate keyword for the text. This tool, which was developed for the JRC by Mike Scott from the University of Liverpool, compares the word frequency statistics of a text with an expected word frequency. The expected word frequency is the average frequency of a word according to a large general-purpose reference corpus. The tool identifies words as keywords if they occur much more often in the new text than they occur in the reference corpus (normalised by the text length). The higher the discrepancy of the word frequency is, the higher is the keyness of the word. For the word frequency table comparison, the tool offers to choose between the *chi-square* and the *log-likelihood* algorithms. The monolingual indexing tool produces as output a list of indexing terms or *keywords* plus their *keyness* (presented as *associates* and their *weights* in Table 1).³

The goal of the training phase is to produce, for each descriptor in each language, long lists of ordinary language words which are statistically related to the descriptor, plus an indication of their strength of association (their weight). The top of the list of ordinary language words for the English Eurovoc descriptor FISHERY MANAGEMENT is shown in Table 1. We refer to the words in the list as *associates* because they are statistically related, or *associated*, words without necessarily belonging to the same part-of-speech or semantic field as the descriptor. Table 1 shows that the associates of the term FISHERY MANAGEMENT are mostly from the semantic fields of 'fishery' ('fishery', 'fish', 'stock', 'fishing', 'conservation', 'migratory', etc.) and 'management' ('management', 'organization', 'agreement', 'arrangement', 'regulation', etc.). However, the list also includes the words 'highly' and 'fly', which are not intuitively linked to either of the two terms of the compound, but which are apparently statisti-

² *IntelliScope Search Enhancer*, version 1.6, by Lernout & Hauspie Speech Products N.V.

³ A tool with a similar functionality is integrated with the software *WordSmith Tools*, developed also by Mike Scott. An evaluation copy of *WordSmith Tools*, which is distributed by Oxford University Press [4], can be downloaded from <http://www.liv.ac.uk/~ms2928/wordsmith> or from <http://www1.oup.co.uk/elt/catalogue/Multimedia/WordSmithTools3.0>.

Table 1 Associates of the Eurovoc descriptor FISHERY MANAGEMENT (Eurovoc code 56410401)

Associate	Weight	Associate	Weight
fishery	2751.07	mediterranean	323.22
fish	1743.80	sea	320.55
stock	1653.37	highly	312.76
fishing	1191.11	session	263.72
conservation	826.47	resource	258.71
management	731.24	arrangement	252.56
vessel	720.05	fly	250.37
flag	533.36	fleet	214.19
organization	525.05	gfc	202.66
agreement	493.99	fisherman	198.93
migratory	424.20	regulation	181.70
subregional	422.25	council	177.69
catch	390.41	...	

cally relevant in the sense that these words occur much more often in the texts indexed with FISHERY MANAGEMENT than in the other texts of our EP collection.

The associate lists are produced by first concatenating all texts which were indexed with a certain descriptor into a meta-document and by then comparing the word frequency list of this meta-document with the word frequency list of the whole EP corpus. The same procedure is carried out for the whole set of descriptors of one language and then for all descriptors for the next language, and so on.

At the end of the training phase, lists of associates like the one in Table 1 will exist for all Eurovoc descriptors in all languages for which training material exists. Currently, the JRC has only produced the English and the German lists of associates for about 3.000 descriptors each.

2.2 Assignment Procedure

When analysing a new text to which Eurovoc descriptors should be assigned, the procedure is inverted. The assignment tool checks for each word of the new text whether it is an associate of one or more descriptors. If the word is an associate, the natural log of the weight (the *keyness*) of the associate is added on to the score of the Eurovoc descriptor for which it is an associate. Many words are associates for more than one descriptor so that the word can add on to the score of various descriptors, with a different weight for each of them.

Once all the associates occurring in the text have been used to produce the cumulative score of the descriptors, the cumulative scores are divided by the number of words of the new text. The result of the assignment procedure is a ranked list of Eurovoc descriptors for the new document, as shown in Table 2.

Table 2 Top 40 Eurovoc descriptors and their score assigned automatically to the policy document *Resolution on linguistic and cultural minority in the European Union*

Score	Descriptor	Score	Descriptor
92	<i>COMMUNITY PROGRAMME</i>	54	ANNUAL REPORT
84	YOUNG PERSON	54	ACTION PROGRAMME
80	<u>CULTURAL POLICY</u>	53	ACCESSION TO THE COMMUNITY
79	CEEC	52	INFORMATION NETWORK
77	<i>EUROPEAN UNION</i>	52	<i>CULTURAL COOPERATION</i>
76	CONTINUING EDUCATION	52	TRANSLATION
68	INTEGRATION INTO EMPLOYMENT	51	STUDENT MOBILITY
66	<u>RIGHTS OF MINORITIES</u>	51	<u>LINGUISTIC GROUP</u>
65	<u>MINORITY LANGUAGE</u>	51	<i>CULTURAL PLURALISM</i>
65	<i>CULTURAL IDENTITY</i>	51	<i>COMMUNITY POLICY</i>
64	<i>EDUCATION POLICY</i>	50	<i>INFORMATION TECHNOLOGY</i>
64	VOCATIONAL TRAINING	49	<i>LANGUAGE TEACHING</i>
64	<i>EDUCATION</i>	48	HUMAN RIGHTS
63	<i>CULTURAL HERITAGE</i>	48	COMMUNITY FINANCIAL INSTRUMENT
63	<i>NEW TECHNOLOGY</i>	47	EXPRUS
61	<u>REGIONAL CULTURE</u>	47	LEONARDO
61	<i>DISSEMINATION OF CULTURE</i>	47	TELECOMMUNICATIONS
59	SOCRATES	47	REGIONAL LANGUAGE
55	<i>MULTILINGUALISM</i>		
55	<i>COMMUNITY ACTION</i>		<u>Underlined:</u> manually assigned descr.
54	EUROPEAN CITIZENSHIP		<i>Italics:</i> further 'reasonable' descr.
54	EFTA COUNTRIES		<u>Strikethrough:</u> obviously wrong

2.3 Discussion of the Keyword Assignment Results

Typically, professional (human) indexers of the European Parliament assign between three and ten Eurovoc descriptors to a text of one or more pages. The number of automatically assigned descriptors is much larger because most words in a text are associated to one descriptor or another, but it is, of course, possible to limit the output of the assignment tool to a certain number (in our example in Table 2, we limited the number to 40), or to fix a threshold for the score (e.g. minimum score is 60). While the archives of the EP need a small number of concise descriptors, the JRC is interested in a larger number of ranked descriptors as document comparison performs better with larger sets of items to compare (see Section 3).

In Table 2, the five Eurovoc descriptors which were assigned manually by the professional indexers of the EP to the European Commission's policy document are underlined. Even though all manually identified descriptors were found by the automatic procedure, they do not rank highest. Instead, they were assigned the positions 3, 8, 9, 16 and 30. However, many other descriptors which were identified automatically are, according to our own judgement, relevant, as well (they are marked in italics). The text clearly is about COMMUNITY PROGRAMMES, CULTURAL IDENTITY and CULTURAL HERITAGE, MULTILINGUALISM, etc. A judgement regarding the question which exact descriptors, and how many, are most appropriate for a certain text is, of course, very difficult and an absolute answer cannot be given.

Most automatically identified descriptors in Table 2 seem to be appropriate, or at least acceptable, but there are some which are clearly wrong, which means that they do not describe the contents of the document. These are EFTA COUNTRIES, ACCESSION TO THE COMMUNITY and CYPRUS (marked by strikethrough). Although three clearly unwanted results out of 40 will not spoil the outcome of automatic clustering of texts and the visualisation of document collections (Sections 3 and 4), the tool should be improved in order to avoid such unwanted assignments.

2.4 Reasons for Unwanted Assignment

The EP mainly discusses political and legal issues and the texts stored in the EP's database therefore do not cover all aspects of life. The list of associates for the Eurovoc descriptor 'Mauritania', for instance, contains many words having to do with fishery. This must be due to the fact that the country Mauritania was mainly discussed in the context of fishery agreements. When assigning Eurovoc descriptors to a text covering the subject area 'fishery', the tool will find many associates pointing to the Eurovoc descriptor MAURITANIA even if the text does not mention this country at all.

We found that many of the obviously unwanted descriptors are geographical terms. As the JRC has an independent tool to recognise geographical references in texts so that the assignment of geographical Eurovoc terms is not necessary, we are considering the exclusion of the assignment of geographical Eurovoc descriptors altogether.

Although the assignment of wrong descriptors is obviously not wanted, their existence may not be harmful for the calculation of cross-language document similarity (see Section 3) because the tool will consistently assign the wrong Eurovoc descriptors in texts of all the languages because the associates found for a descriptor such as MAURITANIA will be biased towards the same subject area in the different languages. The terms of the semantic field of 'fishery' will be dominant in the associate lists of MAURITANIA in all languages.

2.5 Comparing Manual and Automatic Indexing

There is no doubt that manual (human) indexing is of higher quality than automatic indexing. However, human keyword assignment is not perfect either and has its own problems. We were told by the head of a large governmental archive that not only professional indexers usually produce different results from each other for the same text. Their results frequently differ also according to their mood, and they certainly change over time as the view of political and technical issues change. The same document is likely to be indexed with different keywords now than it was indexed a few years ago. Furthermore, it seems that library archives sometimes have to face a high fluctuation of indexers because assigning keywords to a text requires intelligent people with good conceptual skills and knowledge while the daily work of assigning keywords to texts is rather boring. The fact that maintaining a group of human indexers is expensive is barely worth mentioning.

Automatic indexing, on the other hand, is consistent, fast and inexpensive. It can be carried out on the spot for a new collection of documents which, for instance, a

user may just have downloaded from the Internet. Furthermore, the indexing process can be repeated for the complete document archive in case the indexing algorithm is modified, so that consistency is always guaranteed.

3 Calculating Document Similarity Across Languages

Measuring the similarity of documents can be useful if a user wants to find similar documents to the one chosen, or when trying to organise large document collections in order to get an overview of the main structure and contents of the collection. Document similarity measures are usually based on lexical overlap, which means that documents are assumed to be similar if they partially use the same words. The bigger the lexical overlap is, the more similar the documents are. In addition to the mere number of words two documents have in common, the words are usually weighted differently in order to take into consideration how frequent these words are in general and how frequent they are in the specific documents, or other factors.

The JRC's document similarity calculation tool ([5]) does not consider all words of the documents which are to be compared, but it limits itself to using the most significant keywords of the documents, as produced by the keyword identification procedure described in section 2.1. However, instead of using the keyness as produced by the keyword identification tool, the JRC clustering system recalculates the weight of the keywords depending on the set of documents whose similarity is to be computed. The reason for this is that a word like 'fraud' may be a good keyword to describe the contents of a fraud-related document, but within a selection of texts which are all fraud-related, this word is not so useful to calculate document similarity. Keywords which are frequent and which are shared between many of the documents whose similarity is to be calculated are therefore down-weighted.

Figure 1 shows a hierarchical clustering tree graph organising seven English documents according to similarity, produced by the JRC clustering system. The words on the right are the first three of a ranked list of keywords which describe the seven individual documents. The keywords for the clusters of documents are calculated by the

document name	node+attraction	word#1	word#2	word#3	...
agricultural_policy_h.\	53.\	consumer	restoration	encephalopathy	...
consumer_movement_h.../		consumer	labelling	spongiform	...
investment_aid_h...../	43.\	consumer	labelling	transparency	...
community_control_h...../		consumer	spongiform	encephalopathy	...
goat_h.....\	29-\	processing	encephalopathy	spongiform	...
press_h...../		consumer	encephalopathy	bovine	...
cosmetic_product_h...../	22--	monitoring	ban	bovine	...
		bovine	bse	consumer	...
		scrapie	infect	scientific	...
	62.\	scientific	scrapie	veterinary	...
		scientific	bovine	veterinary	...
	42..../	scrapie	scientific	infect	...
		scrapie	encephalopathy	infect	...

Figure 1 Small sample cluster of seven documents and the first three of a ranked list of indexing terms for each document. The JRC clustering system also calculates a ranked list of the most representative indexing terms for each document cluster.

clustering tool, using the keywords of the clustered documents as input. The whole cluster of seven documents in Figure 1 can thus be described by the indexing terms ‘bovine’, ‘BSE’ and ‘consumer’, and by other keywords which could not be displayed here for space reasons.

It is obvious that using lexical overlap (of words or of keywords) as a similarity measure is only reasonable when comparing texts of the same language with each other. If documents of two different languages are to be compared, their contents have to be mapped to a common, more language-independent representation. A list of Eurovoc descriptors which represent the approximate contents of a document is such a language-independent representation which can be used for cross-language document comparison because Eurovoc descriptors are merely numerical codes with translations in each language. This means that assigning Eurovoc descriptors to two documents written in different languages allows the measurement of their similarity. The more Eurovoc descriptors two documents have in common, the more similar they are. The assignment of Eurovoc descriptors to texts provides a unique opportunity of comparing texts in multilingual document collections with each other, of organising them and of producing a ranked list of similar documents to a given document.

4 Visualisation of Multilingual Document Collections

For larger document collections, the organisation in hierarchical clustering trees is unmanageable. The benefit of the organisation into smaller groups of related texts vanishes when the graph spreads over many pages. For this reason, we have worked on a variety of ways to compare documents and to visualise document collections in a different way ([6], [7]).

4.1 Document Maps

A particularly intuitive way of visualising document collections is a document map. Document maps are two-dimensional representations of a document collection, usually built on the basis of a high-dimensional representation of the document space. In these maps, each document is represented by a dot and dots representing similar documents are placed close to each other on the map. The most well-known approach to producing document maps is probably the one by Kohonen et al. [8] at the Helsinki University of Technology who use neural networks both for the similarity measurement between documents and for the organisation of the documents in a two-dimensional space.

Figure 2 shows a map of 258 English documents, which is based on the Eurovoc descriptors assigned to these texts and which was produced with the commercial product *ThemeScape* by Cartia Inc.⁴

In addition to placing dots for similar documents close to each other, the intuitive visualisation of *ThemeScape* uses the landscape metaphor by combining document clusters into islands where clusters formed of a larger number of documents form a

⁴ See <http://www.newsmaps.com> and <http://www.cartia.com>

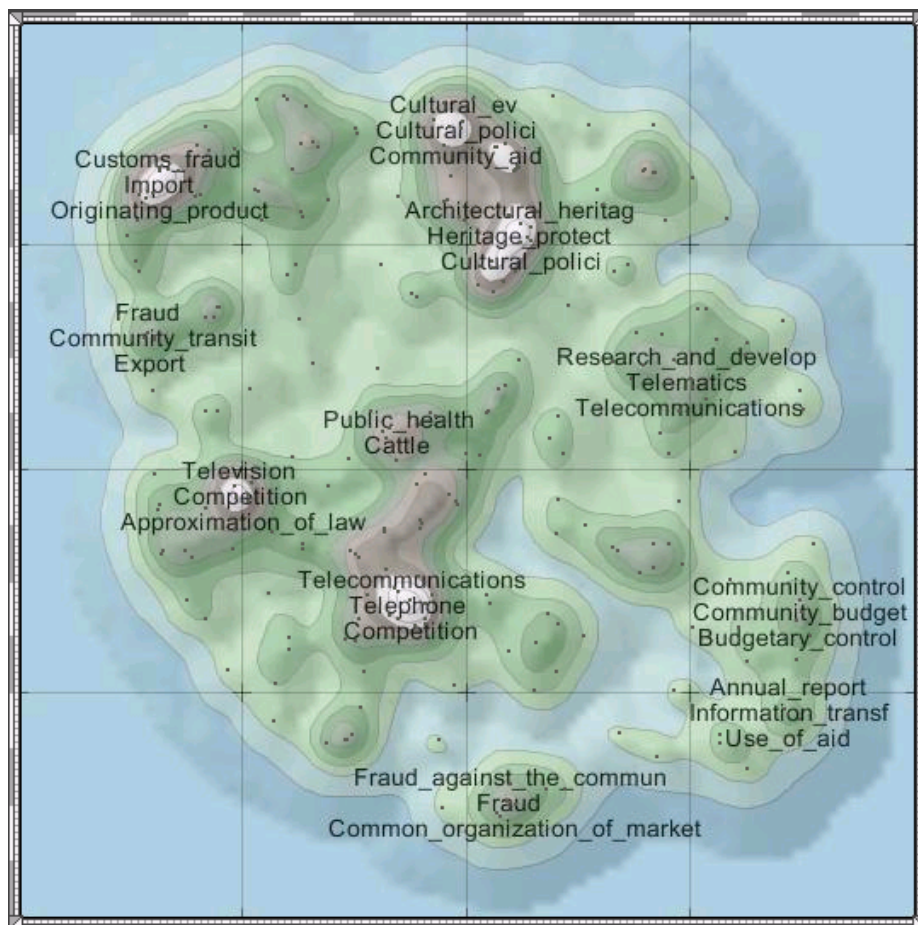


Figure 2 Document Map of 258 documents indexed with Eurovoc descriptors, produced with *ThemeScape*.

hill and the biggest clusters are represented by snow-covered mountains. Additional colours indicate the 'depth of the sea' between the islands in order to guide the eye.

The software gives an overview of the main contents of the document collection through the visual means of the map and by annotating the map with the keywords for each area of the map.

4.2 Representing Multilingual Document Collections in One Map

Cartia's software carries out a document comparison based on lexical overlap. It identifies keywords for clusters, spreads out the documents in two-dimensional space and annotates the cluster islands with keywords. Cartia's approach is monolingual.



Figure 3 Document map showing larger lists of keywords and the lists of documents of a specific area. The map furthermore shows the search results for the word ‘olive’.

The maps can be viewed using any HTML browser equipped with standard functionalities. Users can navigate in the document collection by zooming into areas of interest. Clickable areas allow the user to view longer keyword lists for each area, to display the list of documents of an area and to subsequently read a document (Figure 3). Furthermore users can search for specific keywords and let the system highlight the documents which are described by the searched keyword in the document map.

The JRC has carried out experiments with the *ThemeScope* software by feeding it with lists of Eurovoc descriptors instead of with full texts.⁵ The successful experiment showed that documents written in different languages can be visualised in one single map if a link is provided which allows the calculation of cross-language document similarity.

⁵ Results of this experiment can be viewed at the temporary Cartia web site <http://demo.cartia.com/jrcdescriptors>

5 Outlook

Before handing over a keyword assignment, document clustering and visualisation system for multilingual document collections to users in a working environment, results and procedures have to be optimised and to be evaluated thoroughly. At the current state of affairs, there is still space for improvements and an in-depth evaluation yet has to be carried out. Future plans of the JRC include linking texts to a second language-independent knowledge structure consisting of products and product groups.

5.1 Evaluation of Procedures and Results

Evaluating the automatic assignment of keywords to documents is very difficult because keyword assignment is a conceptual process and there are no clear-cut rules saying that a certain keyword should or should not be assigned. The situation is even worse for the evaluation of document maps. Satisfaction of the users is the only possible measure of their usefulness.

Lacking better alternatives, the comparison of the automatically assigned Eurovoc descriptors with the manually assigned ones seems to be the best solution, even though manual keyword assignment is no guarantee for the creation of error-free and optimal keyword lists (see Section 2.5). Automatic keyword assignment could thus be assumed to be good if the manually assigned keywords rank highest in the list of automatically identified Eurovoc descriptors. This quality measure suggests some steps which can be carried out to improve and to optimise the system.

5.2 Improving the Eurovoc Assignment Results

The most promising ways of improving the Eurovoc descriptor assignment procedure are (a) to compare the automatic and the manual indexing results, (b) to compare the automatic indexing results in documents which are translations of each other, and (c) to use descriptor co-occurrence statistics based on the manual indexing results in order to imitate the manual process as closely as possible and to suppress obvious non-sensual assignments.

According to the approach mentioned as (b), the Eurovoc descriptor assignments to a text and to a good translation of this text should yield identical results because the content of the two documents ought to be the same. The third approach, mentioned under (c), assumes that descriptors which human indexers often assign together to the same text are semantically or otherwise related, whereas descriptors which are rarely or never assigned to the same text are unlikely to be useful combinations. Frequent descriptor combinations in the manually indexed document collection should thus be favoured over combinations which do not occur there. Descriptor co-occurrence statistics based on large amounts of manual indexing results should therefore be useful when tuning the tool.

5.3 Linking Documents to a Second Language-Neutral Representation

The JRC plans to develop an information extraction tool for texts which identifies references to products and product groups from a product nomenclature. A couple of alternative but related product nomenclatures are under consideration. The first one is the *Combined Nomenclature*, which is the European Community's classification of goods for external trade statistics. The second is the *Customs Tariff Code TARIC*⁶. Both nomenclatures have the advantage that translations exist in all eleven official EU languages so that the extraction results of a text in one language can also be viewed in the other ten. Even though the approach to cross-language product reference extraction is similar to the one taken in multilingual keyword assignment, the means will have to be different because there is no training document set which can be used to apply machine learning techniques. Linking a text to a second language-independent knowledge structure should improve cross-language document comparison and hence the visualisation of multilingual document collections.

References

1. Steinberger, R.: Software Solutions to Overcome the Language Barrier. JRC Technical Note No. I.00.91, Ispra, 2000.
2. Steinberger, R. and J. Hagman: Commercial Keyword Identification and Clustering Software. JRC Technical Note No. I.00.90, Ispra, 2000.
3. Thesaurus Eurovoc – Volume 2: Subject-Oriented Version. Edition 3 / English Language. Annex to the index of the Official Journal of the European Communities. Luxembourg, Office for Official Publications of the European Communities, 1995, ISBN 92-77-86366-8
4. Scott, M.: WordSmith Tools v. 3.0, 1999. Oxford University Press, Oxford (UK).
5. Hagman, J.: An Implemented Cluster Analyzer for Documents and their Indexing Terms. JRC Technical Note No. I.00.106, Ispra, 2000.
6. Hagman, J., D. Perrotta, R. Steinberger, A. Varfis: Document Classification and Visualisation to Support the Investigation of Suspected Fraud. In: H. Zaragoza, P. Gallinari and M. Rajman (eds.), *Working Notes of the Workshop on Machine Learning and Textual Information Access at the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'2000)*, 12 pages, Lyon, September 2000.
7. Hagman, J.: Some Ways of Visualizing Results of Cluster Analysis. JRC Technical Note No. I.00.107, Ispra 2000.
8. Kohonen, T., S. Kaski, K. Lagus and T. Honkela. Very large two-level SOM for the browsing of newsgroups. In: C. Von der Malsburg, W. Von Seelen, J.C. Vorbrüggen and B. Sendhoff (eds.), *Proceedings of ICANN'98, International Conference on Artificial Neural Networks*, Lecture Notes in Computer Science, Vol. 1112, pages 269-274, 1006. Springer, Berlin.

⁶ See http://europa.eu.int/comm/taxation_customs/databases/taric_en.htm