Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC

Ralf Steinberger, Bruno Pouliquen, Johan Hagman

European Commission, Joint Research Centre
Institute for the Protection and Security of the Citizen (IPSC)
Cybersecurity and New Technologies for Combating Fraud Unit (CSCF)
21020 Ispra (VA), Italy
{Ralf.Steinberger, Bruno.Pouliquen, Johan.Hagman}@jrc.it
http://www.jrc.it/langtech

Abstract. We are presenting an approach to calculating the semantic similarity of documents written in the same or in different languages. The similarity calculation is achieved by representing the document contents in a language-independent way, using the descriptor terms of the multilingual thesaurus EUROVOC, and by then calculating the distance between these representations. While EUROVOC is a carefully handcrafted knowledge structure, our procedure uses statistical techniques. The method was applied to a collection of 5990 English and Spanish parallel texts and evaluated by measuring the number of times the translation of a given document was identified as the most similar document. The good results showed the feasibility and usefulness of the approach.

1 Introduction

Following the introductory clarification of the question what semantic document similarity is (1.2), why (1.2) and how (1.3) to measure it, and how this work fits in with other activities at the JRC (1.4), section 2 summarises earlier work [11] on assigning controlled vocabulary thesaurus terms (henceforth called *descriptors*) to texts. Section 3 then describes how we use these lists of automatically assigned thesaurus descriptors as kind of a conceptual interlingua which allows to measure the semantic document similarity without using any dictionaries. Sections 4 and 5 discuss the limitations of the adopted method and give an outlook on future work.

1.1 What is document similarity?

Although everybody has an intuition regarding the question whether two documents are similar, and to what extent, it is difficult to put one's finger on this intuition. Similarity measure can be based on the degree of lexical overlap between the texts that are to be compared, but it is also possible to use a more abstract measure by comparing the document contents. Latent semantic indexing approaches, for instance, go beyond counting the mere overlap of words used in texts and map words and documents to a more complex conceptual space [4]. Document similarity can also be based on stylistic information such as sentence length, the type-token ratio, word variation and other stylometric features. Finally, meta-information such as document type,

A. Gelbukh (Ed.): CICLing 2002, LNCS 2276, pp. 415-424, 2002.

author name, source of the text, time of writing and other information aspects could be used.

Our own approach is to map different texts onto an existing knowledge structure, i.e. the multilingual thesaurus EUROVOC (see 2.1 and [1]), which has the advantage that it exists in all eleven official European Union languages. Unlike other approaches to measure cross-language document similarity ([4], [10]), our own approach does not require language pair-specific linguistic data because EUROVOC acts as a conceptual interlingua. On the other hand, our own approach cannot be extended to languages other than the ones covered by the thesaurus used.

1.2 Motivation to calculate the semantic similarity between documents

Who is interested in the automatically calculated semantic similarity between documents written in different languages? What is it good for? Our own motivation for carrying out this work was to help users in the working environment of international organisations such as the European Commission to find their way through large multilingual document collections. One of the functionalities we considered to be useful is the capacity of showing users a ranked list of documents that are similar to one they are interested in, even if these other texts are written in different languages. Another functionality is the one to allow users to navigate through a multilingual document collection, using a document map [2, 12]. Document similarity calculation is also an essential tool for the automatic classification of texts into given classes.

In addition to our own motivation, efforts are made to compile automatically a collection of parallel texts in order to gain statistical knowledge on texts and their translations [5, 10]. Assuming that the translation of a text is the most similar text for a given one, our similarity calculation tool can be used for this application, too.

1.3 How to evaluate automatic similarity calculation

It is a non-trivial question how to judge, even intuitively, document content similarity. Are a three-page text and its 20-line abstract more similar than two 3-page documents talking about a similar subject? Should text length play a role at all in document similarity calculation? And should document language be a factor? It seems intuitively obvious that a text and its high-quality translation should be very similar. However, translators and other people speaking two languages very well know that different languages express concepts differently and have different ambiguities so that it is not reasonable to assume a 100% identity between a text and its translation.

Due to the complexity of the issue, automatic similarity calculation is rather difficult to evaluate. Lacking other alternatives, we decided to use the successful spotting of text translations as an evaluation criterion: we assume that, looking at a large text collection, if our system identifies the translation of a document as the most similar document, it performs well (more on this in sections 3.1 and 3.3).

1.4 The JRC's text management system

The document similarity calculation tool is part of a larger system put together by the European Commission's (EC) *Joint Research Centre* (JRC) that should help to man-

age the *information overflow* and to cross the *language barrier*. The JRC's system has three main components: one component whose task it is to find and retrieve documents in a variety of languages which are potentially relevant for the user's interests [8]; a second component that analyses the retrieved documents and extracts various information aspects from them; and a third component that visualises and presents the textual information and the extracted meta-information in a variety of ways [2, 12].

We consider the tool for calculating the similarity between texts to be part of the second (document analysis) component. The results produced by this tool are required for the information visualisation task, as carried out by the third component.

2 Assignment of EUROVOC thesaurus descriptors to texts

We assign EUROVOC descriptor terms automatically, using a statistical approach that uses a training text collection to which descriptor terms had been assigned manually. For the application presented here, the training corpus consists of 6636 English and Spanish texts from the European Parliament (EP). After the off-line training phase (2.2), the descriptors can be assigned rather quickly online (2.3). Before applying any statistical techniques, we pre-process both the training material and the documents to be indexed by lemmatising all words, marking up the most frequent multi-word terms with underscore (e.g. *human_right*) and defining a large list of stop words. Details of this work have been published recently [11] so that we will only summarise this step here. However, the assignment algorithm has been improved since and has been applied to a text collection other than the set of EP training texts, so that we would like to describe the new algorithm and to present the latest assignment results.

2.1 The EUROVOC thesaurus

EUROVOC was developed by the EP and the European Commission's Publications Office (OPOCE), together with national organisations of the EU member states for usage as a controlled vocabulary to index large multilingual document collections manually. EUROVOC exists in all eleven official European Union languages. Version 3 [1], which we use, consists of 5933 descriptor terms that are hierarchically organised into 21 fields and, at the second level, into 127 micro-thesauri. The maximum depth is eight levels. In addition to the 5877 pairs of *broader terms* (BT) and *narrower terms* (NT), there are 2730 pairs of *related terms* (RT) linking descriptors not related hierarchically. EUROVOC has a wide coverage and contains descriptors from the fields of politics, law, economics, finance, social questions (including culture and religion), education, science, employment, transport, environment, agriculture, forestry and fisheries, foodstuffs, technology and research, energy, geography, organisations and more.

Due to its wide coverage, EUROVOC is useful to describe texts from very different fields, but with less than 6000 descriptors it is not very detailed. Our main reasons to choose this thesaurus over others were that EUROVOC exists in exact translations in all eleven official European Union languages and that we were given access to both the thesaurus and to two manually indexed training collections (one from the EP and one from OPOCE). Furthermore, EUROVOC is used by many national and international organisations so that our work is sure to meet the interest of several user groups.

2.2 Training phase

As EUROVOC descriptors are usually rather long and complex expressions which are unlikely to occur in their exact formulation in the running text of documents, we achieve the assignment of the relevant descriptors by producing automatically, for each descriptor in each language, large lists of semantically and statistically associated words (more precisely: *lemmas*) which, when found in a new text, trigger the assignment of the descriptor. We refer to these associated words as *associates*. For instance, the descriptor #12360607 (English text: PROTECTION OF MINORITIES) has associates such as *racism*, *xenophobia*, *minority*, *protection*, *human_right*, *indigenous_people*, *ethnic_minority*, etc.

We identify these associated lemmas in several steps, exploiting a training collection of texts for which professional indexers from the EP have identified the most appropriate EUROVOC descriptors manually. First, we compile, for each descriptor, a list of all texts of the training collection that were manually indexed with this descriptor. We refer to these text collections as *meta-texts*. We then compare the lemma frequency list of each meta-text with the lemma frequency list of the whole training collection, using the log-likelihood test [3]. The result of this comparison is, for each EUROVOC descriptor in each thesaurus language, a list of key lemmas that are particularly characteristic for this descriptor (associates). In addition to the lemma, a *keyness* value gives information on the degree of relevance of each lemma for this descriptor. This procedure is described in more detail in [11] and [12].

2.3 Assignment phase

During the assignment phase, the lemmas of a new text that is to be indexed with EUROVOC descriptors are compared to the associate lists of all EUROVOC descriptors of the text language. Our assumption is that, the more similar an associate list is to the list of lemmas of the text, the more appropriate the corresponding descriptor is for this text. The descriptors can then be ranked according to their appropriateness, as expressed by an automatically calculated score.

After trying out a variety of different algorithms to compare the text with the associate lists (TFIDF, Cosine, Okapi, and others), we identified the *Cosine* formula [7] as producing the best EUROVOC descriptor assignment results, i.e. the overlap between manually and automatically assigned descriptors was biggest. Experiments showed that a mixed formula, using TFIDF, Okapi and Cosine with varying weights, produces precision results 3% to 6% higher than those shown in Figure 1. However, we did not use this optimised formula because its calculation for new documents is computationally heavier and its results are harder to use for the following document similarity calculation step. Interestingly, the document comparison procedure described in section 3 produces better results when using input (assigned EUROVOC descriptors) produced with the Okapi formula [6]. This shows that, for the purpose of similarity calculation, the *consistency* of the EUROVOC descriptor assignment is more important than its actual precision.

The *Okapi* formula (1) considers the number of times a lemma is used as an associate for a descriptor (DF_l), the number of associates in the associate list of the descriptor ($|d|$), the average number of associates in all associate lists (M), the total number of EUROVOC descriptors (N) and the occurrence frequency of the lemma in the

text ($TF_{l,t}$), according to the following formula, with *d* being the descriptor, *t* being the text and *l* being a lemma (associate).

$$Okapi_{t,d} = \sum_{l \in t \cap d} \log\left(\frac{N - DF_l}{DF_l}\right) \frac{TF_{l,t}}{TF_{l,t} + \frac{|d|}{M}} \tag{1}$$

The *Cosine* formula (3) computes the cosine of the angle of two multi-dimensional vectors [7]. If the vectors are about the same, the angle is about zero, so that the cosine is close to one. In our case, we calculate the cosine of a text's lemma frequency list with the lists of the various EUROVOC descriptor associates and their keyness. The Cosine formula uses the term weighting formula TFIDF (2), with the term frequency $TF_{l,d}$ being the number of times an associate lemma occurs in the meta-text and the document frequency DF_l being the number of descriptors for which the lemma *l* is an associate. So, when DF_l is one (lemma appearing only in this one descriptor), the TFIDF value will be high and when DF_l is about *N* (lemma appearing in all the descriptors), the TFIDF value will be low.

$$TFIDF_{l,d} = TF_{l,d} . ((\log_2 \frac{N}{DF_l}) + 1) \tag{2}$$

$$COSINE(d,t) = \frac{\sum_{l \in d \cap t} TFIDF_{l,d} . TFIDF_{l,t}}{\sqrt{\left(\sum_{l \in d} TFIDF_{l,d}{}^2\right)\left(\sum_{l \in t} TFIDF_{l,t}{}^2\right)}} \tag{3}$$

Figure 1 shows the EUROVOC descriptor assignment results achieved for all the 2432 English texts of our OPOCE collection for which descriptors had been assigned manually. While both the EP and OPOCE use EUROVOC to index the documents in their archives, the two organisations deal with different kinds of texts so that the document collections used for training and for testing are different. In the OPOCE test collection, 5210 different descriptors had been assigned manually (EP training collection: 5142), with an average of 5.21 descriptors per text (EP training collection: 6.59). As our system produces a ranked list of descriptors of user-definable length, precision and recall can be calculated for any number of automatically assigned descriptors. Figure 1 shows that the highest-scoring descriptor (rank 1, x-axis) assigned by our system had also been assigned manually in 53% of all cases (performance on training set: 84%). Had the descriptors been assigned arbitrarily, the success rate for rank 1 would have been 0.088 % (5.21/5933).

As the EUROVOC thesaurus is not a flat list of terms, the relationship between descriptor terms had to be considered in the evaluation. Among the automatically assigned descriptors that had not been chosen manually, those which are an RT, BT or NT to a manually chosen one are *better* results than those which have no recognised relationship with the manually chosen descriptors at all. In addition to the percentage of correctly found manually assigned terms, Figure 1 therefore also shows performance information including RTs, BTs and NTs. While the human indexers were given instructions not to assign both the BT and an NT of a relevant concept, our system exclusively follows the similarity criterion. Figure 1 shows thus that, in 63% of all documents, the highest-ranking automatically assigned descriptor was either manually

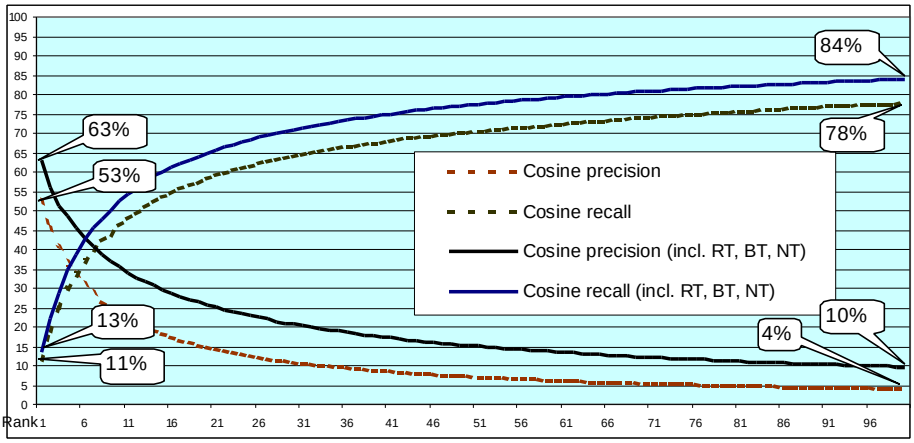

Fig. 1. EUROVOC descriptor assignment results for 2432 English OPOCE documents (Cosine formula), measuring the overlap between automatically and manually assigned descriptors for different ranks.

assigned or it was a BT, NT or RT of a manually assigned descriptor (performance on EP training set: 87%).

3 Document similarity calculation

The ranked lists of EUROVOC descriptors assigned to documents can be seen as an approximative representation of the document contents. Therefore, these descriptor lists can be used to calculate the similarity between documents. The more similar two descriptor lists are, the more similar we expect the two corresponding texts to be.

3.1 Translation spotting vs. similarity calculation

As we mentioned in 1.3, the success rate with which translations of a text are identified as the most similar documents to a given one is the most obvious way of evaluating the similarity calculation performance automatically. The idea is that, within a document collection, the most similar document to a given one should be its translation. However, the task of identifying the translation of a given document is different from finding other similar documents. Firstly, translations are obviously written in a different language from the original text so that the search space is only half the search space of a bilingual document collection. Secondly, translations have a similar length and structure to the original document. These criteria can be used to optimise the performance of the translation spotting exercise. As our Spanish EP training texts used, on average, 13.5% more characters than their English equivalences (length factor LF = 1.135), our translation spotting formula assigned the highest similarity values to those texts that were 13.5% longer and punished texts with a different length in proportion to their deviation (see formula (5) in 3.2).

As the tasks of finding translations and finding other related documents, including those written in the same language and having different length, are different, we carried out two separate experiments: one which generally searches for similar documents (3.2), and one which searches specifically for translations (3.3). For both experiments, we used translation spotting as an evaluation criterion, i.e. the higher the score and ranking of the translation is in the list of automatically identified similar documents, the better are the results. As translation spotting is not our primary concern, but merely an evaluation criterion, we also calculated document similarity without considering the text length and without restricting the search space to the Spanish texts.

3.2 Calculating document similarity based on EUROVOC descriptor lists

We calculated the similarity between documents by calculating the mutual distance between their automatically identified EUROVOC descriptor lists, using a cosine measure [7]. The documents which are the least distant are the ones which are the most similar. The first similarity formula (4) is a cosine on the vector space of the automatically assigned EUROVOC descriptors, with *d1* and *d2* being two documents, *e* being a EUROVOC descriptor, and $score_{e,d}$ being the Cosine or Okapi score of the EUROVOC descriptor for this document. The second formula (5) adds a length factor to the previous one, where *length* is the total number of characters in the document and LF is the language pair-specific length difference (1.135 for Spanish-English; see 3.1). Note that this Cosine formula uses automatically assigned EUROVOC descriptors as input and that these can be calculated with either the Cosine or the Okapi formula (see the discussion in 2.3). The results in Figure 2 are based on EUROVOC descriptors assigned by using the Okapi formula.

$$Sim(d_1, d_2) = \frac{\sum_{e \in d_1 \cap d_2} score_{e,d_1} . score_{e,d_2}}{\sqrt{\left(\sum_{e \in d_1} score_{e,d_1}{}^2\right)\left(\sum_{e \in d_2} score_{e,d_2}{}^2\right)}} \tag{4}$$

$$Simfl(d_1, d_2) = \frac{\min(length_{d1}, length_{d2})}{\max(length_{d1}, length_{d2})} Sim(d_1, d_2) \cdot \mathrm{LF} \tag{5}$$

We used a collection of 2995 English and 2995 Spanish OPOCE texts (total of 5990 texts) that are translations of each other to test the performance of our system. The performance results are shown by the two lower lines in Figure 2. The x-axis shows the rank at which the translation was found, the ideal being that the translation was the highest-ranking document (rank 1; the most similar). We tested for 920 English documents whether their Spanish translation was found among all 5990 English and Spanish documents, and at what rank. Figure 2 shows that in 16% of all cases, the translation was found to be the most similar document (rank 1) to the original English document. Furthermore, it shows that, when using the length factor in the similarity calculation, the criterion of identifying the translation automatically is fulfilled much more successfully (55%). However, according to our own intuition, length should not play a role when solely looking for similar documents. When testing the system on the EP training collection, the results were 30% and 70%, respectively.

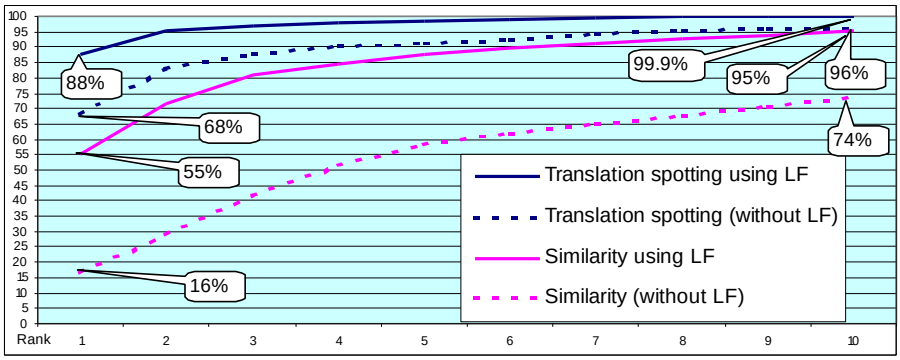

Fig. 2. Performance of the document similarity calculation (3.2) and translation spotting (3.3) tasks, using the automatically assigned descriptors produced with the Okapi formula as input.

The average similarity score of all translation document pairs is about 69% and is thus rather high (s.d. = 0.125; EP training collection: 77%, s.d. = 0.10). This value is slightly lower than the 78% (s.d.=0.09) produced by Landauer and Littman [4]. The fact that the translations *ranked* much lower (while still identified as being 69% similar to the original text) can be explained by the fact that the collection consists of many documents with very similar contents so that these similar documents outperformed the translations. For instance, many texts were resolutions taken by the EP on stopping nuclear tests, with small textual variations depending on the countries they discussed.

3.3 Spotting Spanish translations of English documents

The upper two lines of Figure 2 show the similarity calculation performance when the search space to find Spanish equivalents to English texts is restricted to the 2995 Spanish texts. Again, two different results are given: one for input produced considering the length factor and one for input not considering it. Both are produced using the Okapi EUROVOC descriptor assignment data as input. Applying the length filter again produces considerably better results than when not considering it (88% vs. 68% for rank 1). For the EP training data, the numbers were 93% and 91%.

The translation spotting precision (Spanish translations of English texts found) is much higher than the precision of the similarity calculation task. Presumably, this is not only due to the fact that the search space is halved (only the 2995 Spanish texts were considered as translation candidates). It is likely that our system is also slightly biased towards identifying similar documents in the same language as the original text because the likelihood of assignment of some descriptors may differ from one language to the other.

3.4 Implementation details

The current system is implemented using mainly PERL, CGI and a relational database management system (RDBMS; either Oracle or MySQL) and runs on Unix or Windows/NT. The lemmatiser used is Lernout & Hauspie's *IntelliScope Search Enhan-*

cer. The tool to identify the associate lists for each descriptor is a customised version of the keyword identification functionality of Mike Scott's *WordSmith Tools* [9].

4 Limitations of this method

As with any other automatically trained system, the performance depends heavily on the quantity and quality of the training data. For our current system, we have used training data received from the EP and applied it to texts of a different nature, which we received from OPOCE. As the EP texts do not cover all domains covered by EUROVOC, we do not have enough training data for all EUROVOC descriptors. Furthermore, the sublanguage used in EP texts is rather specific. We therefore expect better coverage (more associates for more descriptors) and better results when adding the OPOCE texts to our training data.

The EUROVOC thesaurus covers a wide range of domains (see 2.1), but it is not very detailed. Mapping document contents to such a relatively coarse knowledge structure means loosing some information when dealing with texts from very specific domains such as highly scientific texts. However, as it is our intention to apply this system to general Commission-related documents and to an automatically gathered collection of online newspaper articles, the detail of EUROVOC should be sufficient.

5 Planned work

Our calculation of document similarity depends on the quality of the EUROVOC descriptor assignment results. We believe that we can achieve better results by improving text normalisation and data cleaning, by experimenting with various parameters, and by using additional training data from OPOCE. Once the process has been optimised for the languages English, Spanish and German, for which the system has currently been trained, we intend to apply it to the remaining EU languages.

The English language knows that "The proof of the pudding is in the eating" so that the ultimate criterion to measure the success of our system will be customer satisfaction. Therefore the application will have to be incorporated in a working system, together with other tools for document gathering, text analysis and information visualisation applications.

References

1. Eurovoc (1995). *Thesaurus Eurovoc - Volume 2: Subject-Oriented Version*. Ed. 3/English Language. Annex to the index of the Official Journal of the EC. Luxembourg, Office for Official Publications of the European Communities. http://europa.eu.int/celex/eurovoc
2. Hagman Johan, Domenico Perrotta, Ralf Steinberger & Aristide Varfis (2000). *Document Classification and Visualisation to Support the Investigation of Suspected Fraud*. Workshop on Machine Learning and Textual Information Access (MLTIA). Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'2000), 12 pages. Lyon, September 2000.

3. Kilgariff, Adam (1996). *Which words are particularly characteristic of a text? A survey of statistical approaches*. Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition, Sussex, April 1996, pp. 33-40.
4. Landauer Thomas & Michael Littman (1991). *A statistical method for language-independent representation of the topical content of text segments*. In Proceedings of the Eleventh International Conference: Expert Systems and Their Applications, volume 8, pp. 77-85, Avignon, France, May 1991.
5. Resnik Philip (1999). *Mining the Web for Bilingual Text*. 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Maryland, June 1999.
6. Robertson, S. E., S. Walker, M. Hancock-Beaulieu & M. Gatford (1994). *Okapi in TREC-3*, Text Retrieval Conference TREC-3, U.S. National Institute of Standards and Technology, Gaithersburg, USA. NIST Special Publication 500-225, pp. 109-126.
7. Salton G. (1989). *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*. Reading, Mass., Addison-Wesley
8. Scheer Stefan, Ralf Steinberger, Giovanni Valerio & Paul Henshaw (2000). *A Methodology to Retrieve, to Manage, to Classify and to Query Open Source Information - Results of the OSILIA Project*. JRC Technical Note No. I.01.016, 35 pages.
9. Scott, Michael (1999). *WordSmith Tools v.3.0*. Oxford University Press, Oxford, UK. www.liv.ac.uk/~ms2928/wordsmith
10. Smith Noah (2001). *Detection of Translational Equivalence*. Unpublished Undergraduate Honours Thesis. University of Maryland, College Park, Maryland, USA.
11. Steinberger Ralf (2001). *Cross-lingual Keyword Assignment*. Proceedings of the XVII Conference of the Spanish Society for Natural Language Processing (SEPLN'2001), Procesamiento del Lenguaje Natural, Revista No. 27, pp. 273-280. Jaén, Spain.
12. Steinberger Ralf, Johan Hagman & Stefan Scheer (2000). *Using Thesauri for Information Extraction and for the Visualisation of Multilingual Document Collections*. Proceedings of the Workshop on Ontologies and Lexical Knowledge Bases (OntoLex'2000), 12 pages. Sozopol, Bulgaria, September 2000.