Zooming in on some Components of a System for Gathering, Analyzing, and Visualizing Multilingual Data

Johan Hagman

Joint Research Centre of the European Commission – Institute for Protection and the Security of the Citizen – Cybersecurity and New Technologies for Combating Fraud – T.P. 361 – 21020 Ispra (VA) – Italy – johan.hagman@jrc.it – www.jrc.it/langtech

Abstract

We are building a system combining the facilities of automatic retrieval of user-relevant multilingual document sets from the Internet; text copy detection; language recognition; keyword assignment; categorization; cluster analysis; and visualization of the results to support querying and data exploration. Knowing that a "chain is never stronger than its weakest link", in this paper we zoom in on some of the modules of this system to discuss their qualities, modus operandi, and various forms of output, depending on the data types and the user's purpose.

Keywords: multilinguality, text comparison, automatic keyword assignment, cluster analysis, data exploration

1. Introduction

The *Joint Research Centre* (JRC) is a department of the *European Commission* (EC) that provides the EC with applied research and services neutral and independent from private and national interests. In one sector[1] within the JRC unit *C bersecurit and New Technologies for Combating Fraud*, language technology (LT) is applied e.g. to support fight against fraud and Internet abuse[2]. This JRC LT group also collaborates with the *European Anti-Fraud Office*, OLAF[3], and its partners in fields relating to systems supporting strategic and operational intelligence. This paper outlines the construction of a general, automated system to help investigators gather and analyze information of interest to current topics, and to present the results in an intelligible way. Special attention will be given to some components of this system.

2. A system for collecting, processing, and presenting textual information

Figure 1 outlines the mission of the LT division within this JRC/IPSC/CSCF/AIM sector. To this aim we are building a system consisting of components corresponding to the processing steps indicated by Figure 1. Some of the components of this system have been described in earlier work (referred to below), whereas here we will zoom in on other modules and aspects.

[1] See www.jrc.it/langtech/ for1 .6(r)-41-5.6t/la -4ro1 .6(rli b)i(1 .6(rn.1(f)s)n)-5 .6(j)-1i.9(c.itt2.1(f)i.9(c.itn.1(f) t.9(c.it2.1(f)is 0 TD

[2] Please, refer to (Scheer et al., 2000; Hagman et al., 2000).

[3] See http://europa.eu.int/comm/dgs/olaf/ for more info about *Office Europ◄en de Lutte Anti-Fraude*.

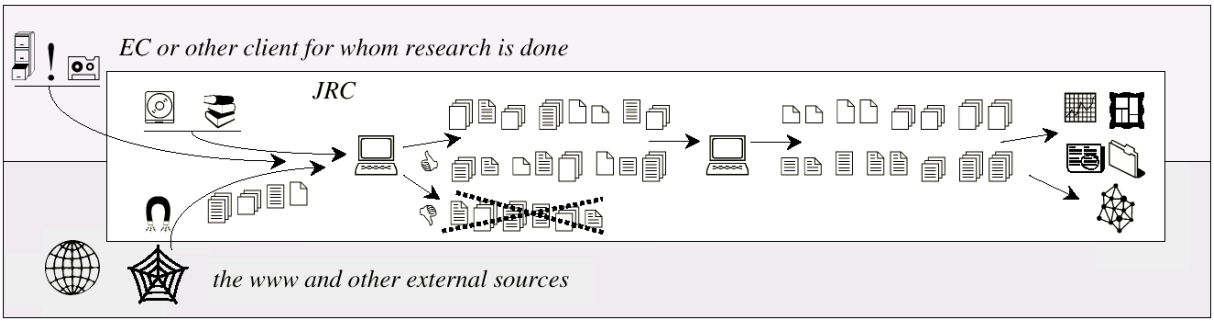


Figure 1. Schematic description of a system for multilingual data gathering, cleaning, language identification, automatic keyword assignment, categorization, and visualization of collected texts

2.1. Receipt of task and allocation of resources

The client for whom we do this *intelligence service* gives us a description of the topic of interest and relevant information already collected by the client himself. We require this information to be in machine-readable format. Having tailored the client's data to our preferred formats, we are able to join or complement them with our own internal information resources, which consists of both EC-internal data and publicly and/or commercially available data.

2.2. Automatic gathering of new relevant information from 'open sources'

Once the description of the client's topic of interest is analyzed, we are ready to instruct our *web crawler* what kind of information it should look for. If we define *manual* gathering of relevant information as the choice we make of what (reference) data to acquire and load into our system (as just described), then *automatic* gathering is that made by this kind of crawler, or *agent*, as it visits the Internet. Note that we use the term 'open source', as it is referred to within the 'intelligence community', i.e. publicly available *information*, not free *source code*. JRC uses both commercially available software and programs developed in-house, all depending on cost, availability, and required degree of customization. We use a web crawler developed and put into service in the exploratory project OSILIA[4]. The crawler visits sites of particular interest, collects the information and puts it in our database of raw data.

2.2.1. Avoiding downloading duplicates

A common phenomenon on 'the Web' is mirrored or duplicate sites, more or less well indicating the original version. As our agent regularly visited a set of newspapers in one application, we also saw the phenomenon of *degrees* of duplication, e.g. articles on evolving events which were being added to and developed through time. We definitely wanted to avoid adding *identical* documents to our database and we therefore applied a filter checking for identical features e.g. the exact file size, and that filtered out some of the text duplicates.

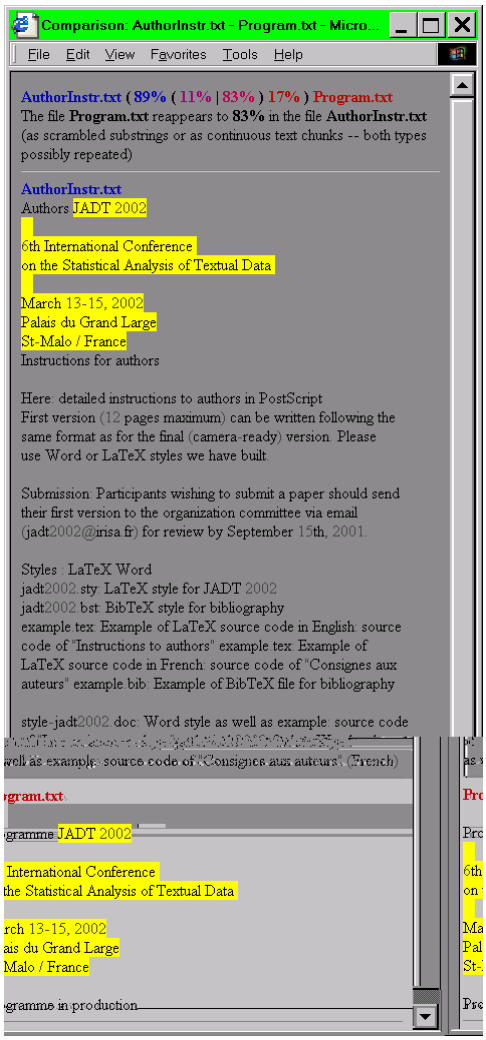
A next step of comparison regards the file content itself, especially the text, lifted out from its HTML embedding. The problem here is how to detect very similar texts and how to stipulate when this *similarity* is to be considered practically equal to *identity*. Just a little piece of "insignificant noise" in one of two otherwise identical files would make it differ from the other file and we may wish to ignore such a tiny difference. On the other hand, sometimes (zooming out a little from the scope of the OSILIA project) one may indeed be interested in

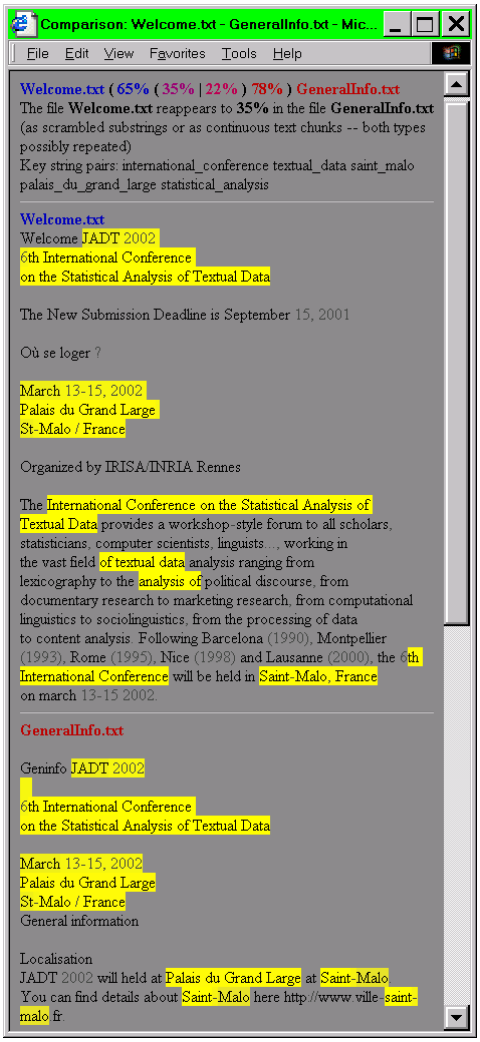
[4] Acronym for *Open Source Intelligence Library on Internet Abuse*, refer to www.jrc.it/langtech/OSILIA.html or (Scheer et al., 2000).

this difference as it could be the signature of somebody pretending to be the author of a text copied from someone else. What is more: two technically different files could be copies of each other, either partial copies, or containing exactly the same text but with the paragraphs presented in a different order. These cases pertain to the issue of *detection of plagiarism*.

We developed a *text comparator* able to detect all consecutive strings contained in both of any two texts, irrespective of where and in which order these strings appear in each text. The amount of shared consecutive word sequences of length ≥2, encountered in any place in the documents, is expressed as an percentage of each document length, respectively, so the value of 100% for one text means that this text *re-appears* or is *included* (possibly scrambled or mixed up) completely in another text. Of course, if two texts relate to each other with this maximum value, it means that they consist of exactly the same text *sequences* (consisting of at least two words) but not necessarily presented in the same *order* in both texts.

As an illustration we take the liberty of running this comparison on the HTML pages announcing this very workshop[5]. Figures 2 and 3 show the result of comparing the clean text versions of the two pairs of HTML pages, i.e. the Author~Instructions ↔ (the~still~undefined)Program (Figure 2), and the pair Welcome~Text ↔ General~Information (Figure 3).





Figures 2 & 3. Example of HTML-formatted output of the document comparator. The percentage of overlap is calculated and shared parts are marked up in ellow

[5] See Appendix. For clarity of this presentation purpose we altered the file names a little.

The first line of the headers in each resulting file is a sort of *Venn diagram* indicating how big a part of each file is contained in their "intersection", i.e. their shared set of word sequences. In fact, only the information 'Program.txt 83 AuthorInstr.txt 11' is passed on to the next module of our system when the option of the HTML-generating output is turned off. The program takes as input a file that defines the alphabet and only those characters are considered in the comparison. Other characters are reprinted (here in gray) but ignored in the process. Certainly one is free to include any characters in one's "alphabet", even numbers and punctuation marks, depending on the type of data to be analyzed. Word sequences occurring in both files are indicated in yellow and the more frequent each word is in both files, taken together, the more the yellow shade tends towards gray. This will – when analyzing bigger files – gray out the function words and leave only the rarer and often semantically "richer words" standing out in bright yellow. A weak tendency of this is seen in Figure 4, which however is still too small a file to dim out these less significant functional words. This statistical effect can also be used in calculating – without any lexicon – possible *ke strings* and *multi-word units* for each given document pair, as indicated in Figure 3.

Running the similarity check on all possible pairs of the seven files in our little example database, we get the similarity indexes shown in Table 1, where we encircle three cases: the files here called 'Welcome' and 'Call4Papers' combine into the largest relative intersection as 88% of the former "re-appears" in the latter. The file 'Registration' has only 7% of its sequences repeated in the file 'Committees', and the same ratio, 7%, it shares with the still unfinished file 'Program'. The part it shares is the same in both relations: it is the name, time and venue of this workshop. As mentioned above, a threshold of a certain percentage of inclusion may be stipulated (e.g. ninety-something percent) where the "more included" file will be ignored and a reference made to the file of which it is the "truest subset".

↑ appears in →	Welcome	Call4Papers	AuthorInstr	Committees	Program	Registration	GeneralInfo
Welcome	100	88	35	34	34	39	35
Call4Papers	32	100	23	11	10	22	16
AuthorInstr	12	31	100	12	11	12	13
Committees	12	13	13	100	12	12	12
Program (Ø)	83	83	83	83	100	83	83
Registration	11	15	9	7	7	100	18
GeneralInfo	22	32	22	18	18	35	100

Table 1. Percentage of shared identical word sequences in seven files

The primary reason for developing this text comparator was to identify and discard duplicates from the set of documents automatically gathered from the web. This was indeed called for since our agent, monitoring some newspaper sites for some weeks, found a ratio of duplicates as high as 25-30% in the collected material. Considering the cases where this is caused by different journalists building on the same texts from the same press service source, we were sometimes able (by activating this visualization functionality) to follow how trees of text editions started growing, observing how the texts were corrected from grammatical errors and complemented with new elements. There were also cases where the same journalist delivered seemingly different articles to different newspapers but which had big chunks in common, detected and visualized by this method.

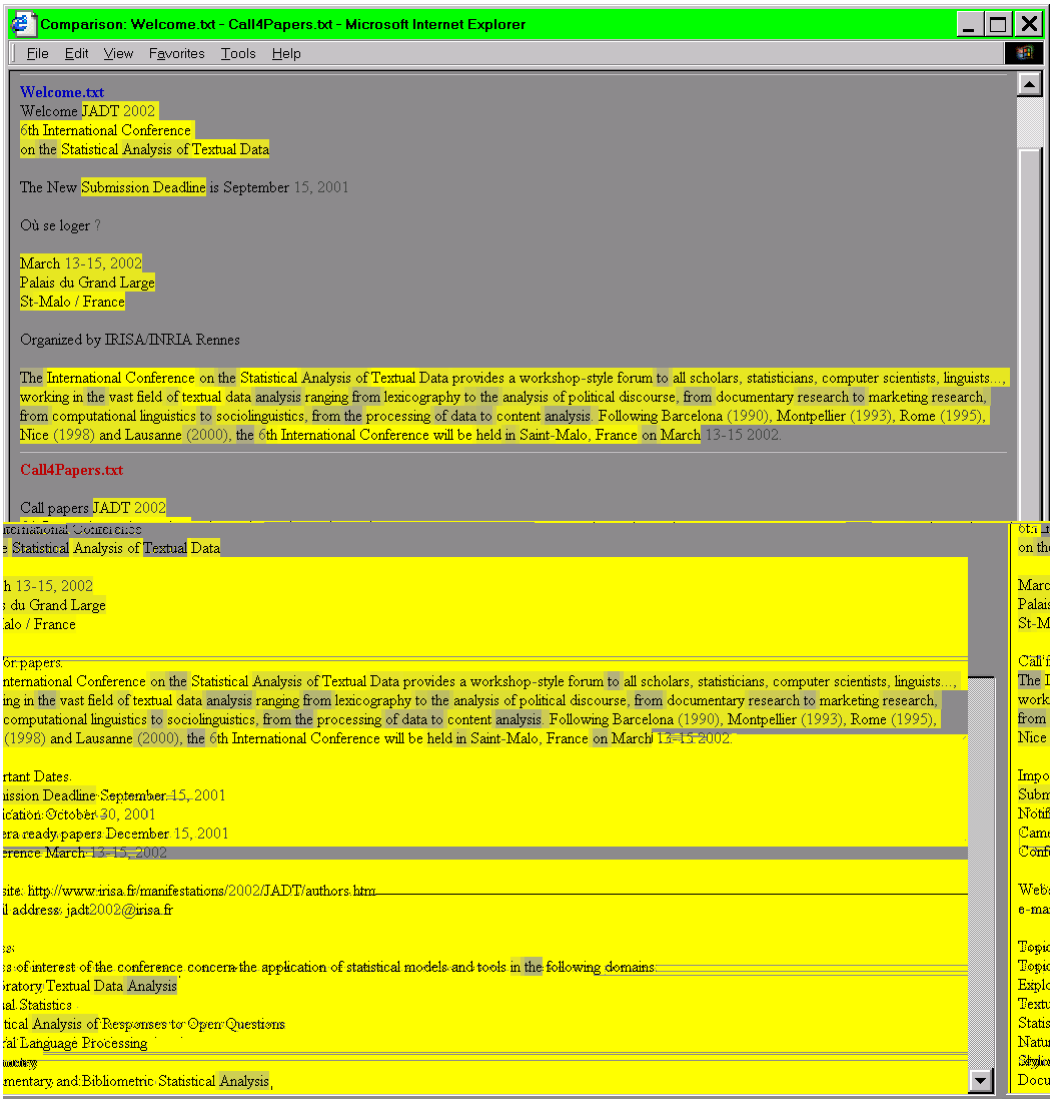


Figure 4. Comparison of the files Welcome and Call4Papers . Sequences at least two words long which are found in both files are marked in shades running from ellow brown gra , depending on the total frequenc of each word in the two files

2.3. Language identification

This stage corresponds to the passage through the first computer icon in Figure 1. A language recognizer[6a] here guesses the predominant language of each text. As this language recognizer (LR) processes the EC text material, it is normally set to decide in which of the EU's ten official languages (using the Latin alphabet, i.e. all except Greek) a document is written. The LR works by comparing the *bigram* sequences in the text with bigram statistics for each of these ten languages. No doubt that using *trigrams*[6b] instead would yield a higher degree of likelihood for each language guessed, but we have found that the somewhat less complex bigram-based algorithm works sufficiently. For each file we retrieve from the Internet we let the LR assign only *one* language. Working with better modularized data than ordinary HTML files, we can even set the LR to assign one language to each *paragraph* or any other discrete *data unit* within that structure – and that is particularly valuable as the EC text material is often a mix-

[6a] See (Hagman, 1999a). [6b] See e.g. (Dunning, 1994).

ture of several languages inside the same document. Let us look at an example of this assignment using the HTML files announcing this workshop last autumn; Figures 5 and 6 show the result of feeding the LR – using two different settings – with the clean text version of that file we here call 'Welcome'.

The whole text in Figures 5 and 6 is considered as one linguistic unit to which a language is to be assigned. The overall bigram statistics suggest *English* as the predominant language and that decides the main text and background colours in this representation. In Figure 5, however, all individual words having another language surpassing English on this basis are marked up with colours suggested by the corresponding national flag. In fact, with the exceptions of 'a', 'to', and 'all', the words guessed to be Romance here (French, Italian, Spanish, or Portuguese) are indeed of Latin origin. Whereas the LR in Figure 5 is set to be "hypersensitive" to indicate non-typical words of the assumed predominant language, in Figure 6 only words that are extremely unlikely to belong to this predominant language are indicated: the un-English '~dt' (in 'JADT') is thought to be Danish and the sequence 'où' is correctly indicated as French. The bigram '~yl~' is so overwhelmingly more Finnish than English so the word 'style' remains frozen stock-still as Finnish.

Committee JADT 2002

6th International Conference
on the Statistical Analysis of Textual Data
March 13-15, 2002
Palais du Grand Large
St-Malo / France

The committees

Program Committee

SPAIN
Ramon Alvarez, Univ. de Leon, Esp.
Monica Bécue, Univ. Polyt.de Catalunya, Esp
FRANCE
Etienne Brunet, Univ. de Nice Sophia Antipolis, Fr
Michel Kerbaol, INSERM, Univ. de Rennes 1, Fr
Dominique Labbe, Univ. de Grenoble, Fr
Ludovic Lebart, CNRS, ENST Paris, Fr (President)
Alain Lelu Univ. de Franche Comte, Fr
Annie Morin, IRISA, Univ. de Rennes 1, Fr
Sylvie Mellet, CNRS, Nice, Fr
Max Reinert, CNRS, Univ. de Versailles SQY, Fr
Andre Salem, Univ. Paris 3, Fr
Pascale Sebillot, IRISA, Univ. Rennes 1, Fr
ITALY
Sergio Bolasco, Univ. de Roma 'La Sap.92 re

As the case with the text comparator described above, when working as a module in our system, for this file, the LR just passes on 'EN 74 FR' to the next module. That value indicates that the text is English with a probability of 74%, having French as it strongest rival. If two languages score close to 50%-50% in number of assigned *whole words*, the number of assigned *single bigrams* is used, trying to tip the scale.

The file 'AuthorInstr', for instance, is assigned the more detailed value 'EN 51 FR / EN 57 FR', saying that 51% of the *words* were summed-up to be English and 57% of the *bigrams* of which they consist are English. The LR was however quite bewildered as it processed the file here called 'Committee'; the result is shown by Figure 7 and the corresponding one-liner result is 'FR 50 NL / EN 51 SV', saying that the file could be written in practically any language out of these four candidates. The user of the system may wish to treat mixed documents like these differently in the following steps of the process and will be notified by this low value.

2.4. Automatic keyword assignment

At this point in the process we should have discarded most of the duplicates and the texts written in a language we are not equipped to process further. The next step is namely *lemmatization* of the text[7]. Once the text is lemmatized, we proceed by assigning various kinds of keywords to each text, as shortly commented in the following subsections.

2.4.1. Named entities

It is often interesting to scan a text for *named entities*. Names of people, geographical locations, companies, products, organisations, and currency expressions may all be interesting indicators of what a text is about. Name recognition software is being offered by a variety of companies and our sector opts for buying such an off-the-shelf tool, as it would be too time-consuming to develop one ourselves. We did however construct our own recognizer of *geographical references*, taking advantage of large lists of geographical place names available from the EC's statistical office EUROSTAT.

2.4.2. Ke words based on natural language

By comparing the *relative* frequency of each lemma in a text with that of a general reference corpus for the same language, we can calculate how *t pical* or *representative* that lemma is for that text; this is expressed by its *ke ness* value. To this aim we use software[8] specially adapted to our needs. By this procedure each text will be given a profile consisting of a list of qualifying key lemmas, their absolute and relative frequency, and their keyness value. Similarities between documents can be calculated based on these profiles and in section 2.5 we will see an example of this. Note that as the non-function words of a natural language typically belong to 'open classes', these keywords constitute a potentially unlimited set.

2.4.3. Ke words taken from a defined thesaurus

The EUROVOC thesaurus was developed by the European Parliament (EP), in collaboration with the EC's Publications Office and several national organizations. The thesaurus exists in exact translations in all eleven official EU languages and covers the major interests of the involved institutions. Hierarchically organized into 21 fields, it contains 127 micro-thesauri with 5,933 descriptor terms altogether. The maximum depth of the hierarchy is 8 levels. One big advantage of being able to assign these descriptors to a text is that they are immediately intelligible in all eleven official EU languages, thereby bridging the language barrier.

---

[7] The lemmatising software used is the *IntelliScope Search Enhancer*, version 2.0, by *Lernout & Hauspie*.

[8] We use a customized version of the keyword identifying functionality of *WordSmith Tools*™, (Scott, 1999).

We have access to a large text collection to which EUROVOC *descriptors* were manually assigned. By calculating the most typical natural language lemmas (in a fashion similar to what is described in 2.4.2) for each text assigned with a certain descriptor, we can establish associative forces (correlations) between natural language lemmas and EUROVOC descriptors and by this method we have achieved quite good results of automatically assigning EUROVOC descriptors to texts which are *reasonabl* semantically similar to the training material[9]. Let us use the HTML files of this conference presentation to illustrate such an experiment as well. We are well aware, though, that the texts announcing this workshop are indeed *not* very similar to the texts coming from the EP's public archive. Instead of a relevance score of about 75 (which we have had in our successful experiments with EP-related material), the scores of these JADT texts rarely reach even 25. We have noticed that texts scoring under 40 are really not useful at all for these experiments but we were still curious to see what would happen when assigning EUROVOC descriptors (trained on completely different text types) to these HTML files.

2.5. Cluster analysis and data visualization

An efficient way to "get the picture" of how the elements in a set relate to each other is to "let them group themselves spontaneously" into clusters. This can be done if their relations (*similarities*, or "vicinities") are based on features expressed in numerical values, which is the case of our texts once provided with keywords and relevance indexes. The cluster analyzer[10a] developed in-house[10b] adopts a hierarchically binary agglomerative algorithm using dynamically adaptive weights for features and subtrees, and clusters either the *items* (here: texts) or their *features* (here: keywords) into dendrograms. There is also a module for 2D cluster projections.

2.5.1. Item dendrograms

Figure 8 shows the tree diagram for the seven HTML files presenting this workshop. The features and similarities of these seven *items* are based on identified natural language lemmas and their respective relative frequencies. The result is not bad. The tree in Figure 9 shows what happens when the files are characterized in terms of what the system guessed to be relevant (EP-debate-trained) EUROVOC descriptors. We remind of the fact that a successful assignment of these descriptors make the result automatically applicable in eleven languages and no translation is necessary from an open set of natural language keywords. In this case, however, we see that trying to recognize EP topics in these texts does not always hit the head of the nail but many of the "guesses" are still fairly good. Some of them are quite amusing: in Figure 9, visiting the leaves in index order (the number in the column where 'Root' appe

```
Call4Papers=\                 1  1 paper work information camera contain reference english lexicography provide processing
          100=\               7  2 forum political ranging lexicography provide processing montpellier contain workshop-style
Welcome=====/  |              0  1 field workshop-style forum linguistics montpellier scientists research marketing
             54=\             9  3 submit document analysis text statistics conference international data jadt organize
AuthorInstr====/  |           2  1 bibliography word french author code postscript final wishing review written detail via pages
                41=\        10  5 jadt text analysis data statistics conference organize saint-malo international palais
Registration\     |  |        5  1 name card form copy cedex register frf written versailles campus georgeault lunches
           83====/  |        8  2 campus register coffee marie-noëlle card georgeault frf student lunches copy form fee breaks
GeneralInfo=/       |        6  1 student marie-noëlle internet georgeault coffee campus general lunches detail website
                  32=\     11  6 irisa/inria organize france jadt text analysis data statistics conference saint-malo
Committees==========/  |     3  1 university rennes committee nice versailles italy france lausanne roma irisa/inria program
                     18= Root  7 palais jadt saint-malo text international france data analysis conference statistics
Program================/     4  1 program palais international saint-malo conference statistics data analysis text jadt france
```

Figure 8. Dendrogram resulting from assigning automaticall natural-language ke words to seven files and then cluster-anal ze their profiles

```
Registration=\                5  1 1011010300000000(approximation_of_laws) 1021030000000000(community_budget)
             |                     0431030104000000(police_checks) 1236020000000000(human_rights) 2841040407000000(health_card)
           100=\               7  2 2841040407000000(health_card) 2446020700000000(quasi-fiscal_charge) 3221012001000000(directory)
             |  |                  6416020203020000(registered_trademark) 6416020203040000(european_trademark)
GeneralInfo==/  |               6  1 3226030404010000(postal_charges) 6411020104000000(harmonization_of_standards)
             |                     1016040200000000(ec_internal_market) 6006080300000000(coffee) 6026010500000000(instant_product)
           28====\             10  3 2831021000000000(literary_and_artistic_property) 2031020300000000(labelling)
             |     |                7241020100000000(mayotte) 7216020300000000(saint_pierre_and_miquelon) 7206040100000000(france)
AuthorInstr=====/     |         2  1 2031010301000000(eco-label) 2811010900000000(community_migration) 4406040900000000(self-employment)
                      |               4411030400000000(wage_earner) 5216040500000000(water_pollution)
                    21= Root  7 7206040100000000(france) 2421021000000000(financial_statistics) 1631030700000000(community_statistics)
                      |               1631030000000000(statistics) 7211030000000000(regions_of_france)
Committees=========\  |         3  1 7206041400000000(belgium) 7206030100000000(spain) 0806030500000000(ratification_of_an_agreement)
                   |  |               1006020500000000(european_parliament) 4426030101000000(strike)
                  22=/        11  4 1631030800000000(economic_statistics) 3231010200000000(information_transfer)
                   |                1631031100000000(official_statistics) 5616050500000000(agricultural_statistics)
Program=========\  |            4  1 3606010000000000(life_sciences) 3216040700000000(student_mobility) 3231030400000000(translation)
                |  |               7206030500000000(malta) 6411040000000000(technology) 9802170900000000(socrates)
               47=/            9  3 6416011200000000(research_and_development) 1021010000000000(community_financing)
                |                6416010100000000(research_programme) 2426010800000000(co-financing) 6411040000000000(technology)
Call4Papers==\  |               1  1 3226050200000000(telecommunications) 3226010100000000(publishing) 6836010101000000(paper)
             |  |                  6836010100000000(pulp_and_paper_industry) 1006020500000000(european_parliament)
            62=/               8  2 2016010000000000(trading_operation) 6416030400000000(applied_research) 3231030100000000(interpreting)
             |                      4411020800000000(employment_statistics) 1631010300000000(economic_indicator)
Welcome======/                 0  1 3221011300000000(report) 2821020200000000(social_indicator) 1016030302030100(cfsp)
                                     2806030101000000(contraception) 6416010800000000(community_research_policy)
```

Figure 9. Dendrogram based on the same text files as those in Figure 8, but the words therein were here used as input to a program which tried to assign EUROVOC descriptors to these files, based on trainig material from EP texts (which were indeed not ver relevant for this t pe of text)

2.5.2. Feature dendrograms

As stated above, even the *features* describing the items are related to each other and we define their internal similarities in terms of co-appearance as descriptors of the items in question. An intuitive measure to the reader might be the correlation coefficient, although we use a more specialized algorithm for our purpose.

Tree diagrams do not have to be *binar* ; the cluster analyzer we use also generates *depend-enc* or *implication* (tree) diagrams and these are particularly interesting when visualizing how features co-occur in a given data set. Figure 10 shows such a diagram containing *natural language ke words* describing EP texts. Note that we cannot use the seven HTML files in our previous example since they are too few to be meaningful when studying keyword/descriptor co-occurrence throughout documents. Figure 10 indicates the term 'fishery' was used as a descriptor for 143 EP texts. In 24 out of these, the term 'mesh' also appeared – and those were all cases when 'mesh' was used altogether. We may say that 'mesh' *depended on* – or *implied the presence of* 'fishery' to 1,000 ‰. The term 'bait' was subordinate to 'mesh' in 10 times out of 10 and that also makes 1,000 ‰ of its total occurrence.

```
fishery       143
 -1000 mesh        24/24
    -1000 bait        10/10
 -1000 shrimp      22/22
 -1000 sardine     17/17
 -1000 undersized  13/13
 - 958 trawl       23/24
 - 944 tac         17/18
    - 818 greenland    9/11
 - 895 senegal     17/19
 - 889 cfp         16/18
 - 864 northwest   19/22
 - 857 nafo        24/28
 - 844 fleet       38/45
 - 814 aquaculture 35/43
 - 625 retain      10/16
```

Figure 10. Part of a stud of how natural-language ke words co-occur in describing a larger EP text corpus. Legend: the ke - word fleet occurs 38 times out of 45 (844) in presence of the term fisher

```
64160101         693       RESEARCH_PROGRAMME
 -1000 36110203      8/8      LINGUISTICS
    -1000 32210115      6/6      MULTILINGUAL_DICTIONARY
 -1000 64160109      5/5      INDUSTRY-RESEARCH_RELATIONS
 -1000 6416010801    3/3      CREST
 -1000 64110106      2/2      ROBOTIZATION
 - 833 6416031      20/24      BASIC_RESEARCH
 - 800 521101050201  4/5      SEA-BED
 - 742 6621020303   23/31      NUCLEAR_FUSION
 - 703 28410403     26/37      HEALTH_SERVICE
 - 700 641103       14/20      ADVANCED_MATERIALS
 - 682 36060405     15/22      OCEANOGRAPHY
    - 700 52110403      7/10      RESOURCES_OF_THE_SEA
 - 667 521102        6/9      GEOPHYSICAL_ENVIRONMENT
 - 615 36060404      8/13      METEOROLOGY
 - 591 5206031001   13/22      NATURAL_HAZARD
 - 566 1006070105   69/122      EAEC_JOINT_RESEARCH_CENTRE
 - 533 64160105      8/15      RESEARCH_BUDGET
 - 506 56060109     39/77      AGRONOMIC_RESEARCH
    - 667 5606010902    2/3      PLANT_BREEDING
 - 485 64160106     16/33      EUREKA
    - 333 3221020101    1/3      SELECTIVE_DISSEMINATION...
 - 333 641601       22/66      RESEARCH_POLICY
 - 222 684604        2/9      CERAMICS
```

Figure 11. Same kind of stud as that shown b Fig. 10 but here the terms are EUROVOC descriptors (i.e. codes + read-out text)

This type of dendrogram reminds us of a thesaurus and – based on a sufficiently large text database – it can be useful when constructing a thesaurus manually or semi-automatically as it would suggest data-derived terms and relations and not only those conceived mentally. It can also be used to assess existing thesauri, e.g. the EUROVOC thesaurus, to study which of all thousands of terms are used at all, how often, and in combination with what other terms, and whether this co-occurrence reflect the hierarchic order of the thesaurus. We did some runs on EP texts (indexed manually by EP staff) and generated the complete inventory of all descriptors ever used, their frequency, and how their presence implied the presence of other descriptors. The result, see Figure 11, was appreciated as it was presented to the office responsible for the development and maintenance of the EUROVOC thesaurus, for the EC in Luxembourg.

3. Concluding remarks

No matter how sophisticated some modules are in the system we construct, if there are even a few poorly performing modules there as well, data quality will suffer and the imperfections will propagate along the subsequent modules and inevitably affect the final result. In this paper we have zoomed in on some processing steps of the system we develop, steps whose importance may easily be overlooked. It is of interest to avoid overloading a document database with duplicates or near-duplicates (letting the user define the threshold), and it is valuable to capture information written in different languages by identifying the language and route the text to the right translator or lemmatizer. The choice of keywords describing a text is crucial, as are the ways these are assigned and weighted as they will constitute the basis of all kinds of similarity measures at later stages. Finally, the choice of algorithm of cluster analysis and method of data visualization are often determining factors of whether the user will understand the results at all and find them useful for the application in question.

Acknowledgement

I work closely together with my colleagues *Ralf Steinberger* and *Bruno Pouliquen*. Each of us concentrates on fine-tuning different modules of the system sketched here, making us interact as tightly as these modules later operate. I would like to thank Ralf and Bruno for providing pre-processed data, which I processed further into the examples referred to in this paper.

References

Dunning, T. (1994). Statistical Identification of Language, *Techn. Rep. CRL MCCS-94-273*, Computing Research Lab, New Mexico State University.

Hagman J (1999a). Construction and Performance of a Language Recognizer. Modus Operandi Deliverable No. 8, *JRC Technical Note No. I.00.108*. 14 pp.

Hagman J. (1999b). An Implemented Cluster Analyzer for Documents and their Indexing Terms, Modus Operandi Deliverable No. 12a, *JRC Technical Note No. I.00.106*. 15 pp.

Hagman J., Perrotta D., Steinberger R., and Varfis A. (2000). Document Classification and Visualisation to Support the Investigation of Suspected Fraud. *Working Notes of the Workshop on Machine Learning and Textual Information Access (MLTIA) at the Fourth European Conference on Principles and Practice of Knowledge Discover in Databases (PKDD 2000)*, 12 pp. Lyon, Sept. 2000.

Kilgariff, A. (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.

Murtagh F. (1985), Multidim'nal Clustering Algorithms, *COMPSTAT Lect s 4*, Physica-Verl., Würzburg

Salton G. (1983). Introduction to Modern Information Retrieval, McGraw-Hill.

Scheer S., Steinberger R., Valerio G., and Henshaw, P. (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 pp.

Scott M. (1999). *WordSmith Tools v.3.0*. Oxford Univ. Press, Oxford, UK. www.liv.ac.uk/~ms2928/wordsmith

Steinberger R., Hagman, J., and Scheer, S. (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 pp. Sozopol, Bulgaria, September 2000.

Steinberger R. (2001). Cross-lingual Keyword Assignment. *Proceedings of the XVII Conference of the Spanish Societ for Natural Language Processing (SEPLN 2001)*, pp 273-280. Jaén, Spain.

Appendix The raw texts of the seven HTML-files referred to in the examples, version of early September, 2001

W 1com JADT 2002 6th International Conference on the Statistical Analysis of Textual Data The New Submission Deadline is September 15, 2001 Où se loger ? March 13-15, 2002 Palais du Grand Large St-Malo / France Organized by IRISA/INRIA Rennes The International Conference on the Statistical Analysis of Textual Data provides a workshop-style forum to all scholars, statisticians, computer scientists, linguists..., working in the vast field of textual data analysis ranging from lexicography to the analysis of political discourse, from documentary research to marketing research, from computational linguistics to sociolinguistics, from the processing of data to content analysis. Following Barcelona (1990), Montpellier (1993), Rome (1995), Nice (1998) and Lausanne (2000), the 6th International Conference will be held in Saint-Malo, France on March 13-15 2002.

Call pap rs JADT 2002 6th International Conference on the Statistical Analysis of Textual Data March 13-15, 2002 Palais du Grand LargeSt-Malo / France Call for papers The International Conference on the Statistical Analysis of Textual Data provides a workshop-style forum to all scholars, statisticians, computer scientists, linguists..., working in the vast field of textual data analysis ranging from lexicography to the analysis of political discourse, from documentary research to marketing research, from computational linguistics to sociolinguistics, from the processing of data to content analysis. Following Barcelona (1990), Montpellier (1993), Rome (1995), Nice (1998) and Lausanne (2000), the 6th International Conference will be held in Saint-Malo, France on March 13-15 2002. Important Dates Submission Deadline September 15, 2001 Notification October 30, 2001Camera ready papers December 15, 2001 Conference March 13-15, 2002 Website: http://www.irisa.fr/manifestations/2002/JADT/authors.htm e-mail address: jadt2002@irisa.fr Topics Topics of interest of the conference concern the application of statistical models and tools in the following domains: Exploratory Textual Data Analysis Textual Statistics Statistical Analysis of Responses to Open Questions Natural Language Processing Stylometry Documentary and Bibliometric Statistical Analysis Textual Classifica-