CAIS/ACSI '93 Information as a Global Commodity - Communication, Processing and Use

21st Annual Conference
Canadian Association for Information Science
Antigonish, Nova Scotia, Canada
July 1993

Multilingual Access to Document Databases

A Steven Pollitt, Geoff Ellis
CeDAR - Centre for Database Access Research
School of Computing and Mathematics
The University of Huddersfield
UK

This paper examines the reasons why approaches to facilitate document retrieval which apply AI (Artificial Intelligence) or Expert Systems techniques, relying on so-called "natural language" query statements from the end-user will result in sub-optimal solutions. It does so by reflecting on the nature of language and the fundamental problems in document retrieval. Support is given to the work of thesaurus builders and indexers with illustrations of how their work may be utilised in a generally applicable computer-based document retrieval system using Multilingual MenUSE software. The EuroMenUSE interface providing multilingual document access to EPOQUE, the European Parliament's Online Query System is described.

Natural language

As the Moghul emperor Akbar Khan is reported to have discovered (Blakemore 1977), following an experiment to find "natural language" whereby twelve babies were interned for twelve years with dumb nurses, there is no "natural" spoken language. Not surprisingly, to most of us today, the children failed to develop speech, instead they communicated their thoughts by gesture, the absence of a spoken language evidently didn't deny them an ability to think!

Our language is the product of hearing, seeing and, above all, using words. Each individual has a vocabulary specific to them where each word or phrase will evoke a unique meaning. New words or phrases may be given some meaning after a dictionary look-up where other words may link the new word to existing meaning. Building this vocabulary is hard work and is a major component of our formal and informal education.

We have all witnessed the frustration for an individual when asked to define a word and then to define the words of the definition. There is after all an individual understanding of the meaning of a word brought about by using it for a purpose that one hopes will be shared in communication with others. Sharing this understanding and resolving different meanings is just one of the problems we need to overcome in our everyday existence. The jealousy with which some people guard their own meanings of words is one of the major obstacles to shared understanding. The difficulties we experience within a language are much worse between languages, especially when the languages are from very different cultures.

The following dramatic illustration of misunderstanding illustrates the cost of our failures in communication and understanding:

> Because of a mix-up in translation of a single word (mokusatsu) in their reply to the Potsdam ultimatum of the Allies in July, 1945, the Japanese Cabinet was understood to have **ignored** the ultimatum, contrary to their real intention of **reserving comment** on it, with all the dire aftermath of Hiroshima and Nagasaki that we know.
>
> Jagjit Singh quoting Stuart Chase's book The Power of Words
> in Great Ideas in Information Theory, Language and Cybernetics (1966)

The language barriers in the way of the effective workings of the European Community, with nine official and numerous unofficial languages, are obvious. The scope for misunderstanding is immense. Not surprisingly, translation is one of the major activities in the European Community (EC) Institutions both in meetings between delegates from the member countries and for the production of official documents. The EC Institutions, in particular the European Parliament, provide the backdrop to this paper given the political imperative to provide systems which can facilitate multilingual document access to progress the development of the Community.

Computerised language translation systems have yet to provide good quality translations without significant translator involvement. A review of the major efforts at translation would help place the task of those working to translate natural language queries in perspective, however the nature of the user query for document access and an analysis of the fundamental problems of information retrieval suggest that these activities should be distinguished one from another. Those seeking to use machine translation techniques to improve document access must recognise the lack of success in those areas which have received significant attention and investment, and realise that the translation of a document is a very different task to the interpretation of a query that is intended to access that document.

The Role of Indexing Languages, Automatic Indexing and Thesauri

> ... an index record will simplify those attributes of a document it is describing, making the search of the index easier by reducing both the amount of material to be searched and the complexity of the search process, thereby reducing the time required to perform a search....the use of this intermediate communication channel requires the use of a language or coding that is comprehensible to all parties to the communication, in this case the searcher, the cataloguer or indexer, and, in a sense, the documents being searched for (because the index is supposed to represent the reality of document content)
>
> Charles T Meadow in The Analysis of Information Systems 2nd Edition (1973) p.17

There is a rich inheritance of ways in which the searching of documents can be facilitated. It is important that these are given appropriate attention in the research and development currently underway. There may be a tendency to focus on new technology, such as parallel processing, when it is more appropriate to apply existing knowledge on information management which can produce significant improvements for the searcher.

The increasing storage and processing power of the computer has encouraged approaches to document access which concentrate on alternatives to human indexing. The legacy of some of the earliest experimentation is still to a large extent influencing the thinking of researchers.

The following conclusion to a review of the Cranfield tests of the 1950's and early 1960's, was written by Karen Sparck-Jones, in a text that, even twelve years after it was published, should be compulsory reading for those embarking on research into improving document access.

> ...it has proved very difficult to undermine the major result of Cleverdon's work, namely that indexing languages, including natural language, tend to perform much the same.
>
> Karen Sparck-Jones in Information Retrieval Experiment (1981) p.283

This conclusion has been taken by some to indicate that the expense in generating and applying an index language, by either automatic or manual means, cannot be justified. However the provision of simple free-text searching cannot be regarded as sufficient if the performance of the retrieval system is to be at all respectable, as determined by Blair & Maron (Blair and Maron 1985). Others, notably Gerard Salton with SMART, have sought to enhance the performance of retrieval systems and worked hard to build automatic indexing and retrieval systems which reflect the principles of human indexing by generating term associations and term hierarchies.

The demonstration of improvement in retrieval performance and its extensive reporting were not sufficient, however, to result in the general uptake of these automatic procedures, as Salton himself remarks:

> Those results turned out to have little immediate impact on operational information retrieval, largely because of the difficulty of rendering believable test results obtained with sample collections of a few hundred documents when the operational environments include several million items. Additional problems are posed by the enormous investments already made in the available commercial systems, which make it impossible to contemplate a complete retooling of the kind involved in introducing language analysis methods based on the availability of document abstracts and new file organisation methods.
>
> Gerard Salton in Information Retrieval Experiment (1981) p.322

It may be some consolation to Salton that some of these techniques have since been adopted as enhancements to conventional retrieval systems (e.g. the Zoom facility on the European Space Agency Host system, ESA-IRS). It is interesting to note that Salton also applied his ideas to multilingual information retrieval some twenty years ago (Salton 1973).

The economics of building and maintaining document access systems in the 1990's may well see a re-emergence of efforts to bring about effective automatic indexing. The European Parliament has invited tenders for demonstrations of automatic indexing for the EPOQUE (European Parliament Online Query System) database. Unlike the SMART experiments, this automatic indexing is to be employed in a conventional inverted file form of retrieval system, whereby the indexing process will seek to add terms to the stored document references from the EUROVOC controlled vocabulary thesaurus.

The role of the thesaurus in improving retrieval performance is not disputed by those seeking more automatic techniques. The benefits are clearly described by Dagobert Soergel (Soergel 1985).

> By browsing through an index language that is displayed in a well-designed classified structure, a user can clarify and focus on her own image or concept of her need. The structure of the index language serves as a catalyst to crystallise the need. This results in a better query formulation and better retrieval results.
>
> Dagobert Soergel in Organizing Information: Principles of Data Base and Retrieval Systems (1985)p. 239

The Promise of Intelligent Systems for Document Access

The prospects for improving access to documents by applying techniques which have been developed in the course of research into Artificial Intelligence were seen to be good, where the skills and knowledge of the search intermediary could be coded into the system.

Many millions of dollars (and ECUs) have been spent in building computer systems for document retrieval which have been described as intelligent. Indeed, one important component of the IMPACT (Information Market Policy Actions) programme, established by the European Commission in the late 1980's, sought applications to improve access to information across the European Community through the building of "intelligent front-ends".

Before the first IMPACT programme call for proposals in 1988 a demonstration of "an expert systems approach to document retrieval", CANSEARCH, had been built and described at conferences (Pollitt 1985, 1986a) and subsequently in the literature (Pollitt 1987). The approach was promising but was limited to a specific subject domain (Cancer Therapy).

Using expert system techniques to emulate a human intermediary seemed like a good idea at the time, even though the evaluation showed that its performance didn't match the level of a human intermediary for the majority of the test queries, it did outperform a human intermediary on numerous occasions. Other example systems tackled different domains, PLEXUS for plant diseases (Vickery et al 1987) and EP-X (Krawczak et al 1985) for environmental pollution. An excellent review of this area is provided by Ralph Alberico and Mary Micco (1990).

PLEXUS provided the basis for a commercial product, Tome Searcher marketed by Tome Associates, which was used in the 2.7 million (Canadian) dollar MITI (Intelligent Multilingual Interface Systems to European Databases) project funded by the EC IMPACT programme (Blake 1992). Tome Associates disbanded in the course of the project which was started in 1990 and originally scheduled for completion in March 1992. A prototype, due by October 1992, was intended to provide access to some 20 databases hosted on four different systems. The outcome has not been fully reported but it is unlikely that MITI will lead the way forward to improving multilingual access to documents, as there is no continuation into the next stage of the IMPACT programme, similarly it is not being presented as an exemplar demonstration in the IMPACT dissemination of results (European Commission DG XIII/E 1993).

It is not so much that expert systems cannot be applied with some success (Bramer 1990), as many applications are now in daily use, having been absorbed into the techniques regularly applied to the solution of an information processing need. This is particularly true where the purpose of the system has not been confused by an inappropriate simulation of a person to person dialogue (Pollitt 1986b). It is more the case that the domain of document retrieval presents difficulties that the existing techniques - and future techniques cannot overcome. These difficulties concern our use of language and the limited bandwidth for query expression.

Smeaton (Smeaton 1992) suggests that Natural Language Processing (NLP) has become more robust, reliable and efficient in recent years and makes out a case for the use of NLP for automatic indexing and for Conceptual Information Retrieval where the system is performing more of a question answering function than retrieving relevant documents. Unfortunately he fails to appreciate the difference between a natural language query and the natural language of a document.

> If a user has a vague information need then it can be expressed even imprecisely as a statement in natural language or as a Boolean combination of keywords. The user thus requires access to information which itself has been encoded imprecisely in an ambiguous language, natural language.
>
> Alan F Smeaton, Progress in the Application of Natural Language Processing to Information Retrieval Tasks in The Computer Journal (1992) p.268

Smeaton is, on the face of it, advocating a system with **double imprecision.** This need not be the case. The users query can be general but precise using an interaction with a structured vocabulary and the selection of high level concepts (see below). The user may not provide a query statement but simply follow the vocabulary hierarchies to explore the database. Using NLP to create an index for a document is more appropriate given that the sentences in a document indexed using NLP will presumably benefit from being in the context of other sentences. This will assist in resolving ambiguity and take advantage of an extensive vocabulary to ensure indexing takes place at as specific a level as possible, securing good precision and recall on retrieval.

It is interesting to note that Alberico and Micco (Alberico and Micco 1990) see the incorporation of controlled vocabularies as crucial to the successful application of expert systems in information retrieval.

> Standardization and improvement in subject access to recorded information worldwide must be accomplished before expert systems can be truly useful for information retrieval. When that happens, it will become possible to implement expert systems that support vocabulary control and/or mapping natural language keyed by the reader onto terms used in the information system.
>
> Ralph Alberico and Mary Micco
> in Expert Systems for Reference and Information Retrieval (1990) p.185

Accepting the problems of natural language the use of vocabulary mapping from natural language input can be seen to provide a sub-optimal system for query expression to the search system.

Criticisms by John Searle (Searle 1984) of the claims for Artificial Intelligence (AI) have provided an interesting parallel to what is now being proposed as the basis for multilingual document access. Irrespective of whether the philosophical disputes regarding AI will resolve anything, the example of the non-Chinese speaker locked in a room with several boxes of Chinese symbols, able to answer questions in Chinese by exchanging symbols with the outside world, has provided a clue as to what we can achieve with a simpler approach to the problem. This approach is taken in Multilingual MenUSE (described below) and promises to be both more effective and more applicable than systems adopting the natural language approach, making more direct use of the intelligence of the user in appreciating and applying controlled vocabulary.

Multilingual MenUSE

The promise of intelligent interfaces has struck the rock of applicability, somewhat like the efforts of Gerard Salton, except where the aims of the interface have not been particularly ambitious, such as in coping with the different command languages used on database hosts.

The problem of subject based intelligent interfaces are to do with general applicability, having demonstrated their usefulness in narrow domains. This problem was confronted when the rule-based principles underpinning the design of CANSEARCH could not readily be exercised in tackling a different subject area, Biochemical Genetics. Searching using a more pragmatic approach resulted in MenUSE (Menu-based User Search Engine) which, with none of the search statement construction rules embedding the knowledge of the intermediary in the system, could have general applicability (Pollitt 1988, Blake 1992).

The intelligence is seen to be in the user and attempts to emulate intelligent behaviour in the system are discarded. The difficulties in achieving effective document access are now being tackled applying user intelligence with a well-designed interface which facilitated access to the structured vocabulary as suggested by Soergel.

The translation to provide multilingual access to documents is now concentrated on the thesaurus and is a matter of switching displays to the required language. There is no command language, as such, the query is performed by selection. This achieves an independence from the keyboard which makes this approach especially attractive for languages with non-roman scripts such as Japanese.

The example of applying the principles of MenUSE in a multilingual context is EuroMenUSE, searching the European Parliaments Online Query System in all official Community languages (and Japanese) with EUROVOC, the terminology of the EC Institutions.

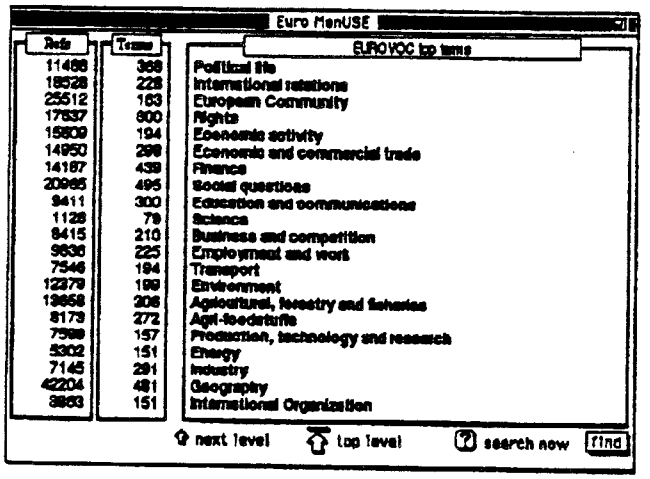
EuroMenUSE

Central to EuroMenUSE is the multilingual EUROVOC thesaurus (OOPEC 1987, 1992), first published in 1984 and used to index the Official Journal and the document references on the European Parliament's document database and CATEL, the production database of the Office for Official Publications of the EC. There are some 5,800 preferred terms structured into 21 fields and translated into all nine official languages (Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish). In addition to the European Parliament, EUROVOC is used in the Spanish Congress, the Assembly of the Portuguese republic, the Turkish Parliament and in the three languages of the Belgian Chamber of Representatives.

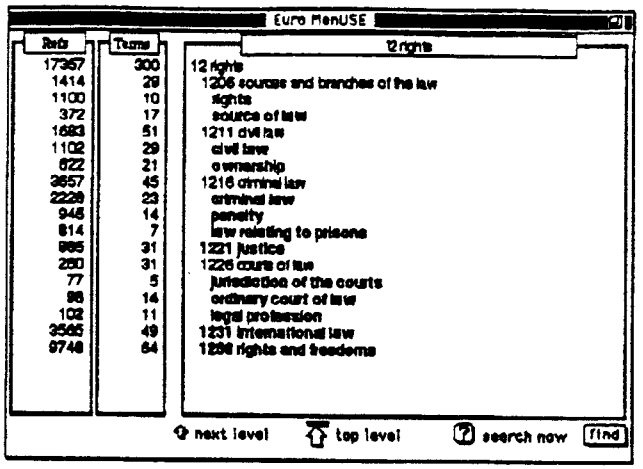
A stated intention is for EUROVOC to provide a common indexing language for different EC databases by acting as:

> ..a bridge linking all the documentation systems relating to the activities of the EC, both within the Community Institutions and at national and regional level and even within the private sector"
>
> EUROVOC Thesaurus User Manual First Edition January 1992.

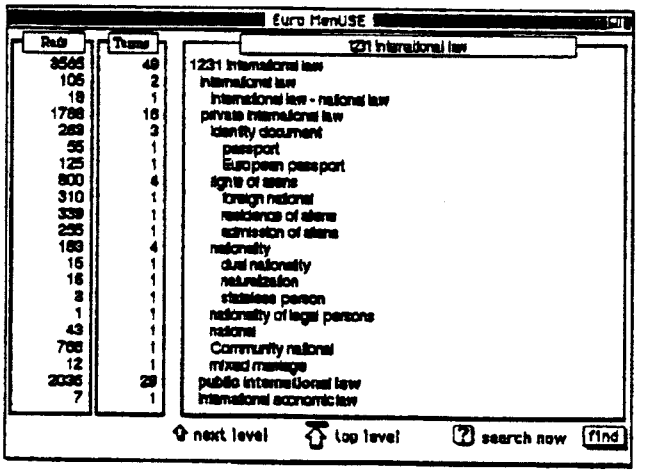
EuroMenUSE processes the complete set of preferred terms against the EPOQUE database so that the structured presentation informs the user how many documents (the Refs column) relate to given concepts at each level of the term hierarchy . The terms column shows the number of terms in that section of the term hierarchy. The top level menu gives an effective overview of the topics dealt with by the European Parliament.



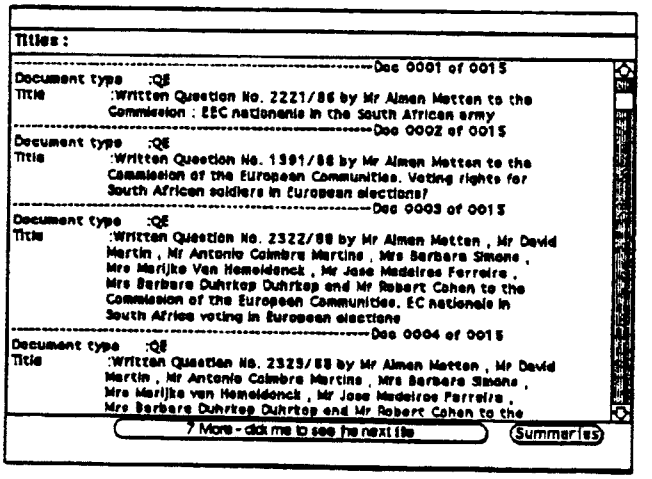
The user navigates the menu by mouse and selects to see a more detailed menu expanding the field on **Rights.** The next menu breaks down the 17,537 documents into the different groups of terms. The automatic procedure which builds menus expands the terms below the second level terms displayed until no further expansion is possible within the physical limits of the screen.



This top down access continues with the user selecting **international law** to review the 3,565 documents broken down using 49 more specific terms.

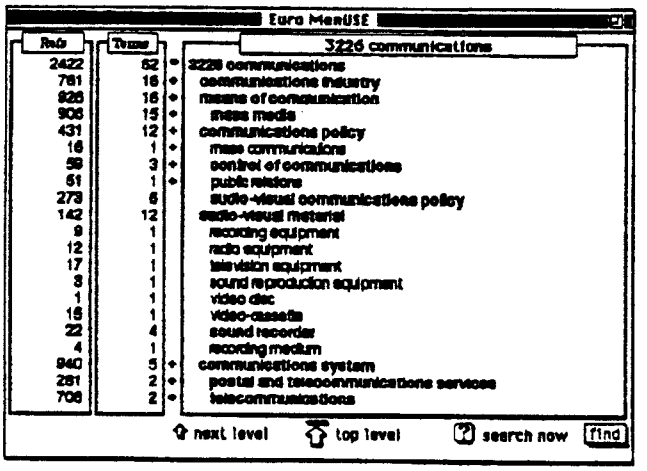


It may be that the set of documents displayed against a term is small enough to be viewed, or that the user wants to get some indication of the kind of documents in a set. The user elects to see the documents indexed by dual nationality.



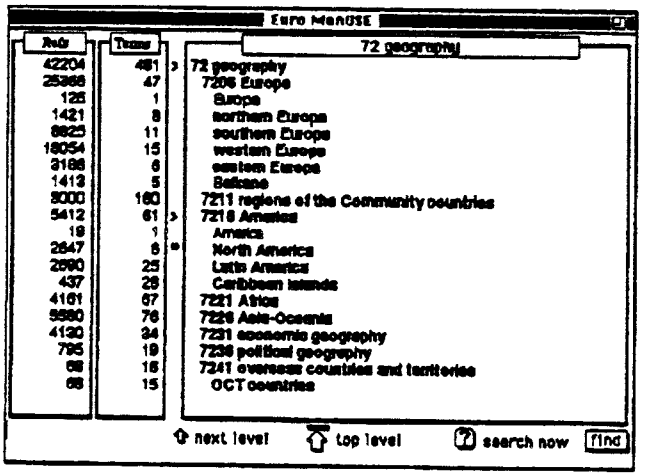
It is possible for the user to enter a term directly. The menu containing that term (or its synonym) is then presented in the context of its broader and narrower terms. This assists the user in identifying the meaning of the term in the vocabulary and enables the user to specify more accurately the subject of their enquiry.

Identifying sets of documents to perform Boolean searching is straightforward. The user can select all 62 Communications terms and then deselect concepts or terms as appropriate. It is impossible to match this efficiency of expression in natural language. Selections and deselections are indicated by markers.

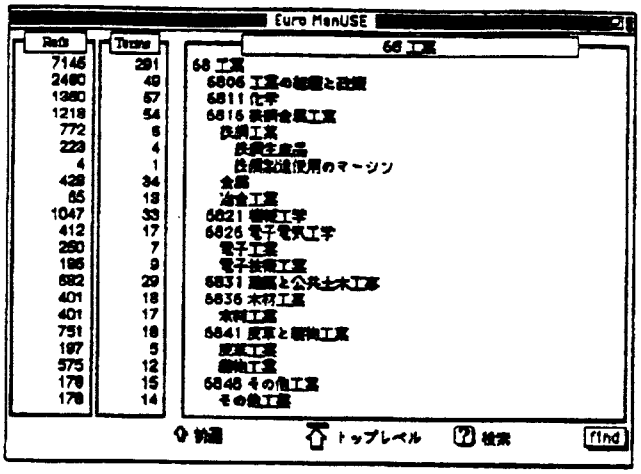


This selection can be incorporated in a Boolean query with other selections.

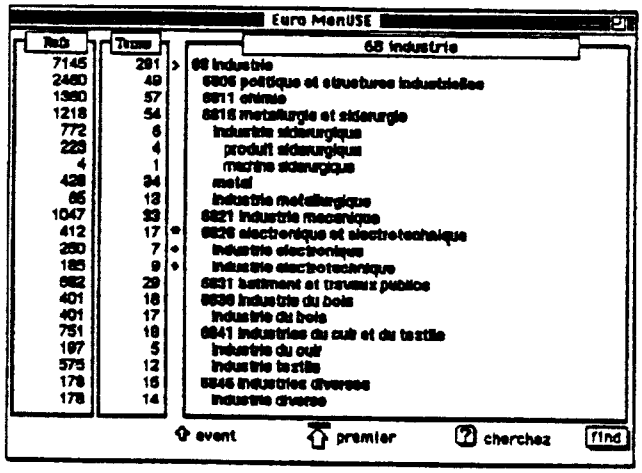
To demonstrate how this can be performed we first select all of North America from the Geography menu.



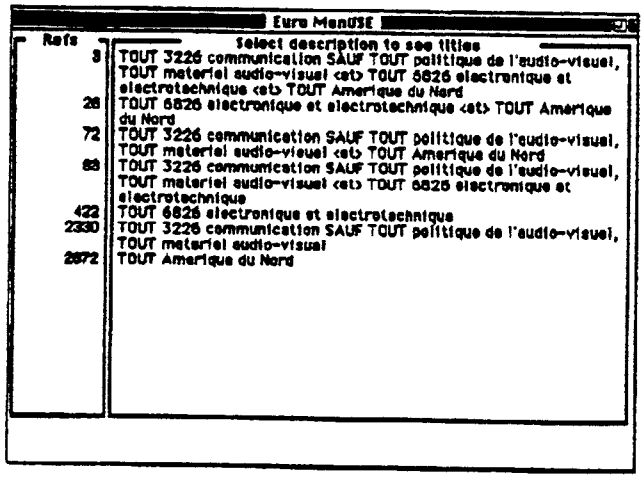
To illustrate the multilingual capabilities we can switch to Japanese and select a breakdown of the Industry top level term.



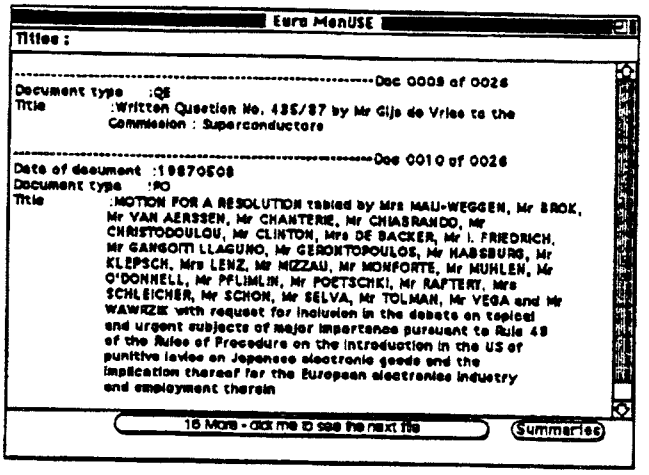
We then switch to French before we select Electronics and Electrical Engineering.



A search is now requested which automatically combines the three selections. The results from the Boolean searching (seven combinations) are presented as a list of set descriptions sorted in ascending order of the number of documents retrieved.

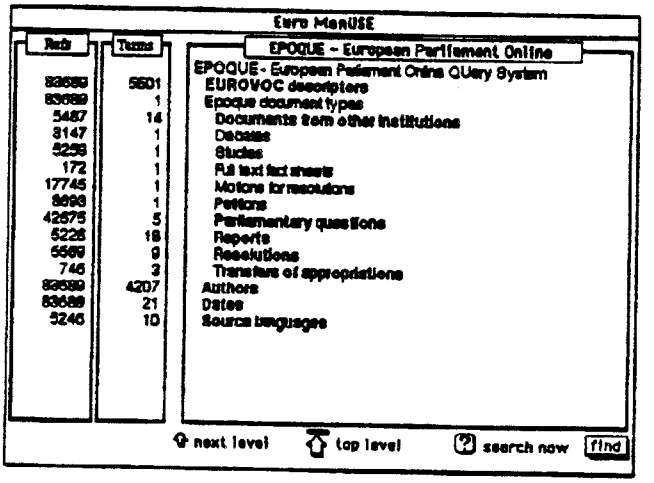


Selecting a summary description will retrieve the document references which are in this example given in English, but should be available on EPOQUE in all official EC languages.



The performance of the system relies heavily on user acceptance of the thesaurus. Certainly it is difficult to find agreement on the structure of a multilingual thesaurus, given that each culture has a view of the world which the thesaurus will reflect. In subsequent versions of EUROVOC perhaps these different views can be accommodated in different structures. In the meantime work is underway to find an acceptable European view for EUROVOC which is acceptable to its multinational user community.

The above example has concentrated on using the EUROVOC indexing terms for searching but for an effective system the attributes from all searchable fields in the document record must be made selectable. The top level menu for EuroMenUSE indicates the possibilities for specifying report type, author and date as additional search parameters.



Conclusion

What we see in MenUSE is a search expression amplifier and qualifier, giving additional power of description over natural language querying. This is coupled to a dynamic classification filter which identifies sets of documents in a database against different combinations of the components of the search expression. Moreover the interaction can be in the language of the user, independent of the language of the documents.

The limited bandwidth in our communications in either speech or writing has received a boost from what might simply be termed a brain extension. Whilst brain scientists are actively engaged in a study of how the brain works in detail, it is apparent from observation and experience that the functioning of our brain is being mirrored in our use of classification and thesaurus structures. Arthur Koestler describes the spotlight and abstractive forms of memory. The former provides for the recall of "past episodes or scenes, or details of scenes, with almost hallucinatory vividness", the latter is structured for retrieval:

> ... if stored knowledge and experience are to be retrievable (for otherwise they would be useless), they must be ordered like a thesaurus or a library subject-catalogue, with headings and sub-headings but also with a wealth of cross-references to assist the process of retrieval....Without hierarchic order and classification, memory would be bedlam...
>
> Arthur Koestler on Memory in Bricks to Babel 1980.

Later in the same chapter, Koestler describes our ability to extend these hierarchies by learning from experience enabling us to be more specific in our descriptions and more discriminating "..Thus we learn to abstract finer and finer nuances - to make the trees of the hierarchies of perception grow new shoots...". He also describes the experience which provides a model for comparison with our indexing activity, whether achieved by manual or automatic means.

> The recall of an experience would then be made possible by the co-operation of several interlocking hierarchies, which may include different sense modalities, for instance, sight and sound or odour, or different branches within the same modality. You may remember the words of the aria "Your Tiny Hand is Frozen", but have lost the tune. Or you may remember the tune after having forgotten the words. And you may recognise the unique timbre of Caruso's voice on a gramophone record, regardless of the words or the tune he is singing. But if two, or all three of these features have been abstracted and stored, the recall of the original experience will have more dimensions and be the more complete.
>
> Arthur Koestler on Memory in Bricks to Babel 1980

Multilingual MenUSE has demonstrated a way to open up access to documents by taking more advantage of user intelligence than alternative "intelligent interface" approaches, overcoming the limiting mechanisms of natural language search specification. These principles are taken further through the simultaneous presentation of the summarised attributes of stored documents in HIBROWSE (Pollitt et al 1993), with the promise of an even richer interaction with the document database.

Acknowledgements

We are grateful to the senior management at the University of Huddersfield who have supported this work and to our colleagues at the Centre for Database Access Research, Martin Smith, Chun Sheng Li, Ann Jones and Mark Gregory who assisted in the development of Multilingual MenUSE. A special thanks is due to Carloyn Watters who encouraged the production of this paper.

References

Alberico R, Micco M (1990) *Expert Systems for Reference and Information Retrieval.* Meckler, Westport 1990.

Blair D C, Maron M E (1985) *An evaluation of retrieval effectiveness for a full-text document retrieval system.* Communications of the ACM 28 (3), 289-299.

Blake P (1992) *The MenUSE system for multilingual assisted access to online databases.* Online Review Vol 16 No 3 pp 139-146.

Blakemore C (1977) *Mechanics of the Mind.* Cambridge University Press, Cambridge 1977.

Bramer M (ed) (1990) *Practical Experience in Building Expert Systems.* John Wiley and Sons, Chichester 1990.

European Commission DG XIII/E (1993) *IMPACT Information Day.* Luxembourg, January 1993.

Koestler A (1980) *Memory* in *Bricks to Babel*. Hutchinson, London 1980

Krawczak D A, Smith P J, Shute S J, Chignell M (1985) *EP-X: A knowledge-based system to aid in searches of the environmental pollution literature.* Proceedings of the 2nd Conference on AI applications, Miami Beach, Dec 1985, IEEE pp 552-557

Meadow C T (1973) *The Analysis of Information Systems* 2nd Edition. Melville Publishing Company, Los Angeles 1973.

OOPEC (1987) *EUROVOC Subject-oriented thesaurus* Vol 2. Edition 2 ISBN 92-77-25830-6

OOPEC (1992) *EUROVOC Thesaurus User Manual* First Edition January 1992 ISBN 92-823-0371-3

Pollitt A S (1985) *AI and information retrieval : tackling some of the problems of naive end-user search specification and formulation* ARTINT, A workshop on Artificial Intelligence and Information Retrieval CEC, Jean Monet Building, Luxembourg, September 1985

Pollitt A S (1986a) *CANSEARCH: An expert systems approach to document retrieval* 10th Cranfield Conference on Mechanised Information Transfer Cranfield Institute of Technology, July 1986

Pollitt A S (1986b) *Reducing complexity by rejecting the consultation model as a basis for the design of expert systems.* Expert Systems, Vol. 3 no. 4, pp 234-238, October 1986

Pollitt A S (1987) *CANSEARCH: An expert systems approach to document retrieval* Information Processing and Management, Vol. 23, no 2, pp 119-138, 1987

Pollitt A S (1988) *A common query interface using MenUSE - A Menu-based User Search Engine.* 12th International Online Information Meeting, Vol. 2, pp 445-457, London, Dec. 1988

Pollitt A S, Ellis G P, Smith M P and Li C S (1993) *HIBROWSE - Adding the power of relational databases to the traditional IR architecture - the future for Graphic User Interfaces.* 15th BCS IRSG Research Colloquium on Information Retrieval, Strathclyde

University, March 1993.

Salton G (1973) *Experiments in multilingual information retrieval.* Information Processing Letters 2, 6-11 (1973).

Salton G (1981) *The Smart environment for retrieval system evaluation* in *Information Retrieval Experiment*, K Sparck-Jones (Ed), Butterworths, London 1981.

Searle J (1984) *Minds, Brains and Science: The 1984 Reith Lectures.* British Broadcasting Corporation, London 1984.

Singh J (1966) *Great Ideas in Information Theory, Language and Cybernetics.* Constable, London 1966.

Smeaton A F (1992) *Progress in the Application of Natural Language Processing to Information Retrieval Tasks*, The Computer Journal 35(3) 268-278.

Soergel D (1985) *Organizing Information: Principles of Data Base and Retrieval Systems*, Academic Press, Orlando 1985.

Sparck-Jones K (1981) *The Cranfield Tests* in *Information Retrieval Experiment* K Sparck-Jones (Ed), Butterworths, London 1981.

Vickery A, Brooks H, Robinson B, Vickery B (1987) *A reference and retrieval system using expert system techniques.* Journal of Documentation, Vol 43 No 1 pp 1-23