Elaine Svenonius
Graduate School of Library and Information Science, UCLA

Compatibility of Retrieval Languages Introduction to a Forum

Svenonius, E.: Compatibility of retrieval languages. Introduction to a Forum.
In: Int. Classif. 10 (1983) No. 1, p. 2–4, 19 refs.

The article, meant to serve as the introduction to the papers presented at a conference on the topic of compatibility between indexing and retrieval languages held at Columbus, Ohio, Oct. 17, 1982, is the proposal that led to the mounting of the conference. It identifies the question to be addressed at the conference, gives a brief overview of compatibility efforts and highlights present concerns. (Author)

1. The Question

The question to be addressed at the conference is: to what extent are two retrieval languages compatible? Considered in the context of multiple data bases, this question is equivalent to establishing the correspondence among thesauri or classification schemes and among sets of query types. By generalizing slightly, the question can be reframed: to what extent are two patterns of language usage compatible? Considered in the context of the interaction between a user and a textual data base, the generalized question is equivalent to establishing the correspondence between a query and the text passages that constitute a relevant response.

The question is interesting intellectually in that answers to it must ultimately incorporate a linguistic theory. But the question has significant practical import as well. One consequence of the "information explosion" is a loss of control that results from the proliferating of data bases. Control can be lost when several data bases have to be searched to find all relevant information bearing on one subject. The problem is particularly acute when the data bases to be searched are in different natural languages, since this necessitates multiple translations of a user's search prescription. But even within the confines of a single natural language, the problem is serious enough. It can happen for instance that several related data bases contain information on a sought subject and access to each is provided through a slightly different command language and a slightly different index language. In this case as well multiple translations of the user's search prescription are needed, the difference being the translation takes place between two retrieval vocabularies within the same natural language.

If manually performed, the translation from one retrieval vocabulary to another is an onerous task in terms of both time and money consumed. It is also one that is likely to be imperfectly performed if it relies on only the semantic relations perceived by the searcher at the time of search. The ideal situation would be to have a search prescription formulated in a user's own (natural) words automatically translated into each of the vocabularies of the various data bases. The extent to which such an ideal situation is achievable is the extent to which an "intermediary" can be dispensed with in online searching. Possibly it is also the extent to which searching may be regarded as algorithmic, rather than intellectual or creative in nature.

2. History

Indexing theorists in the early 1960s were concerned with problems of convertibility and compatibility. At the time the two words were differentiated. "Convertibility" was a term applied to vocabularies and was regarded as "the ability to go from one indexing vocabulary to another"(1). "Compatibility" applied to information systems was defined as "the ability of one information system to accept the original indexing and abstracting data of another information system for any given subject coverage that is common to both systems" (2). Generally, attempts to achieve compatibility adopted one of two methods: 1) constructing a switching language or 2) mapping or directly translating from the vocabulary of one index language to that of another.

A switching language is a mediating language that takes as input the vocabulary of one index language and switches it into the vocabularies of (several) other index languages. The first switching language, so-called, was the "Intermediate Lexicon" developed by the Groupe d'Etude sur l'Information Scientifique in Marseilles (3). The other method of achieving compatibility is simply to translate or map one index language to another and to perform this translation for every pair of languages where it is desired. This method requires the construction of a conversion table. An early example of such a conversion table in this country was the "Dictionary of Equivalents" that was developed to link the vocabularies of the Armed Services Technical Information Agency and the Atomic Energy Commission (4).

During the late 60s and early 70s interest in compatibility problems waned. Not surprisingly perhaps this waning was paralleled by eclipsed interest in machine translation (MT) research. The last few years, however, have seen a reawakening of interest in both MT and compatibility research. Among the forces that have brought about this reawakening are the following:

1) increased availability of inexpensive computing power;

2) Improved understanding, resulting in better consensus, among computational linguists in ways to approach syntactic/semantic problems in automated language processing;

3) in Europe, gaining momentum in cooperative efforts to achieve control across language barriers;

4) in North America, an ever-more-pressing need to achieve control across data bases that index and abstract related disciplines; also, in the United States, an incipient desire to develop an integrated subject authority file for nationwide library use;

5) greater national and international monetary support of MT and compatibility projects.

The present mood toward MT has been characterized as one of "quiet optimism" that "should not be lightly dismissed" (5). The same might be said of research into the compatibility of retrieval languages.

3. Present Research and Development

One indication of the surfacing of interest in compatibility questions is that within the last few years several international conferences have addressed the topic. The most impressive-looking of these took place in Latvia, 6–8 September 1977. The conference was entitled "Unified System of Information Retrieval Languages" (6) and hosted no less than 54 papers on the subject. Included were papers with titles like: "Compatibility Among Classificatory Retrieval Languages", "Measuring the Compatibility Level of Document Retrieval Languages" and "Some Design Concepts and the Structure of a Compatible Language System of a Republic Computer-based STI System".

Three other international conferences at which compatibility concerns were discussed were: 1) "Third Congress of Information Systems and Networks: Overcoming the Language Barrier" organized by the Commission of the European Communities and held in Luxembourg 7 May 1977 (7); 2) "Translating and the Computer", organized by the Technical Translation and Informatics Groups of Aslib and held in London 14 November 1978 (8); and 3) "Natural Language in Information Science", organized by the Committee on Linguistics in Documentation of the International Federation for Documentation and KVAL Institute for Information Science and held in Biskops-Arnö, Sweden, 3–5 May 1976 (9).

In Europe today practical attempts to achieve retrieval language compatibility for the most part focus on the construction of multi-lingual thesauri; to some lesser extent there is still interest in the construction of intermediate switching languages. Perhaps the most well-known instance of the latter is the Broad System of Ordering. Developed for UNISIST, its primary function is to "serve as a switching mechanism to link different individual classifications and thesauri in the process of information transfer" (10), but it has other functions as well, for instance, to specify the subject coverage of serials in a world register of serials and to filter requests for information as they are presented to a UNISIST referral network. Another current example of a switching language is the Intermediate Lexicon for Information Science developed at the North London School of Librarianship by Verina Horsnell et al. The project at the North London School is noteworthy in being one of the few in which tests have been performed to evaluate the effectiveness of a switching language in mediating index languages with varying vocabularies and structural characteristics.

Instances of vocabulary conversion in the form of multi-lingual thesauri abound in Europe where language barriers are a matter of daily inconvenience. A project that has received much recent attention is Eurodicautom, a computerized terminology system designed for users of Euronet (11). It is being developed under the auspices of the European Commission as a part of its three year action plan "Bringing Order out of Babel". Another widely publicized project, on-going at the British Library, has as its aim to make PRECIS translingual (12). In France at the Institut Textile de France, TITUS II, a sophisticated documentary retrieval language, incorporating a normalized syntax as well as a multilingual thesaurus, has been constructed for use in the mechanical translation of document abstracts (13).

On this side of the Atlantic, at the Canadian National Library a subject authority file project is underway to link the Repertoire de vedettes matiere, the Canadian List of English Subject Headings and the Library of Congress Subject Headings (14). While some small-scale multilingual thesauri projects are operational in the United States, most of the current compatibility effort is focused on automated subject switching mechanisms for searching multiple data bases in a single natural language. One example is the R.T. Niehoff project at Battelle Laboratories, the objective of which was "to identify, compile and integrate existing energy vocabularies from systems, both government and nongovernment, into a common indexing and retrieval guide" (15). The project produced a conversion guide, designed for automated, online, switching, that recognized exact, singular-plural and synonym equivalences. Another noteworthy project is that of R.S. Marcus at the Massachusetts Institute of Technology. His conversion mechanisms, in addition to providing a common command language, CONIT, to link DIALOG, MEDLINE, Intrex and ORBIT, also incorporate a partially controlled list of the vocabulary terms used to index the different data bases (16). At the University of Illinois, M.E. Williams developed a mapping model to link chemical data bases. She is presently engaged in a feasibility study to examine "transparent systems", i.e. systems that convert "the procedures, conventions and terminology of one system into equivalent procedures, conventions and terminologies of other systems" (17).

A somewhat different project, but one that is related at its theoretical base in the authority control project sponsored by the Council of Library Resources and being conducted by the Research Libraries Group (RLG) together with the Washington Library Network (18). The aim of this project is to develop an "integrated consistent authority file service for nationwide use". Present efforts are restricted to developing name authority records, but future plans involve addressing the issue of subject authority control. The issue might well be addressed by attempting to merge various subject authority files, e.g. those of the New York Public Library, the Washington Library Network, Stanford and the University of Chicago.

The above examples of retrieval-language compatibility projects are illustrative of the work going forward in Europe and the United States. Some of this work is ambitiously conceived and should it have a successful outcome it could contribute significantly to achieving better data base control, nationally and internationally. There may be some problems, however.

4. Some Problems

While various methods to achieve retrieval language compatibility have been developed, there have been few demonstrations of effectiveness and few cost justifications. This is unfortunate and could lead to wasted ef-

fort as was earlier the case with mechanical translation. That there is need for caution is given credence by remembering that compatibility is a form of vocabulary control, and that vocabulary control, having been examined in both experimental and operational situations, always seems to prove less effective than anticipated. One could reasonably argue that the testing situations were poorly designd; nevertheless, the doubt is there and it is all the more disturbing in view of the high costs associated with the construction and maintenance of vocabulary control devices. We need well-designed evaluation studies, and we need them soon.

Another cause for concern is the lack of theoretical underpinnings in some of the ongoing compatibility projects. Some of the theoretical problems in need of attention are linguistic in nature. For one, while switching languages have generally relied on a broad classification, there is an opposing body of opinion that holds that a switching language must necessarily use a vocabulary more specific than that of any of the languages to be translated (19). This vocabulary may be so specific, in fact, that the feasibility of a true switching language may be conditional upon the possibility of finding a method of semantic factoring. How true this is and whether semantic factoring is a will-o'-the-wisp kind of goal are questions we need to investigate.

Another theoretical problem is that the construction of conversion tables for multilingual thesauri and automated switching mechanisms requires operational criteria of when two terms, or two phrases, are 'equivalent'. Although degrees or levels of equivalences are usually recognized, they tend not to be well defined; nor, perhaps, are they sufficiently numerous. There is a particular difficulty with generic equivalences. Different index languages operate at different levels of specificity. Attempts made to adjust these levels inevitably result in some information loss; on the other hand if no attempt at adjustment is made, then only partial conversion can be achieved. The amount of information lost and the degree to which conversion can be achieved would seem to be concepts susceptible of measurement and thus treatment as variables in empirical investigations. We need feasibility studies.

A third theoretical problem arises because of the special disambiguation problems that attend the conversion of one retrieval language to another. Language, when it is used for retrieval, that is, for indexing or searching, represents a special use of language, a special "language game". One characteristic that distinguishes this use of language from other uses is that generally it utilizes a simpler syntax; one effect of this is a greater potentiality for ambiguity. For instance, in a uniterm retrieval language, there is virtually no syntax; thus, the meaning attaching to individual uniterms is restricted to what they can sustain independent of any context. When translating between two natural language texts, syntactic information can be used to resolve ambiguities; but this approach cannot be used when converting between two uniterm vocabularies. The example given is extreme. There are many different index languages, incorporating differing amounts of syntactic information. The degree of syntactic information incorporated by an index language again may be something that lends itself to measurement. In any case, we need to examine the relationship between conversion feasibility and retrieval language syntax – how much there is and how it is expressed, whether by natural language means or by some nonstandard means such as faceting.

The above problems are only illustrative of directions for research into retrieval language compatibility. What other problems are there and how serious are they? What approaches to solution may be taken? Such are the questions that might be addressed at the proposed conference.

The need for caution mentioned at the beginning of this section should not dampen the present mood of quiet optimism, but should rather contribute to it maturity and judgment.

References:

(1) Henderson, M.M. et al.: Cooperation, Convertibility and Compatibility among Information Systems: A Literature Review. National Bureau of Standards Miscellaneous Publication 276. Washington, D.C.: Government Printing Office. 1966. p 78.

(2) Hammond, William: Dimensions in Compatibility. In: Information Systems Compatibility, edited by Simon M. Newman. Washington, D.C. Spartan Books, 1965. p. 7.

(3) Coates, E.J.: Switching Languages for Indexing. Journal of Documentation 26 (1970) p. 103.

(4) Henderson, op. cit. pp. 78ff.

(5) Hutchins, W.J.: Machine Translation and Machine-aided Translation. Journal of Documentation 34 (1978) p. 150.

(6) Edinajasistema informacionno-poickovych jasykov. (Unified system of information retrieval languages: Summaries of papers presented at an All-Union Conference held in Yurmala, Latvian SSR, 6–8 Sept. 1977). (In Russian). Riga: Latv. Universitet, 1977.

(7) Overcoming the Language Barrier. (Third Congress of Information Systems and Networks organized by the Commission of the European Communities and held in Luxembourg 7 May 1977.) 2 vols. Munich: Commission of the European Communities, 1977.

(8) Translating and the Computer, edited by Barbara M. Snell. Amsterdam: North-Holland, 1979.

(9) Natural Language in Information Science, edited by D.W. Walker, H. Karlgren a. M. Kay. Stockholm: Skriptor, 1977.

(10) Development of a Broad System of Ordering for UNISIST Purposes. UNISIST Newsletter No. 2 (1973) p. 3.

(11) Goetschalckx, J.: Eurodicautom. In: Translating and the Computer, edited by Barbara M. Snell. Amsterdam: North-Holland, 1979. pp. 71–76.

(12) PRECIS for Multilingual Use. Int. Classif. 3(1976) p. 36.

(13) Ducrot, J.M.: The TITUS II System. Boulogne sur Seine: Institut Textile de France, 1974.

(14) Durrance, D.J.: Subject Authority Control in the Canadian Context. Paper presented at the Technical Services Coordinating Group Workshop, Canadian Library Association, Halifax, 1976.

(15) Niehoff, R.T.: Development of an Integrated Energy Vocabulary: Final Report, April 1975 to Feb. 1976. Columbus: Battelle Columbus Laboratories, 1976. (NTIS: PB 253 781.)

(16) Marcus, R.S. and Reintjes, F. Francis: Computer Interfaces for User Access to Heterogeneous Information-Retrieval Systems. Cambridge, Mass.: Mass. Inst. of Technol. 1977. (Report ESL-R-739.)

(17) Williams, Martha E.: Online Retrieval – Today and Tomorrow. Online Review 2(1978) p. 353–366.

(18) Council on Library Resources: An Integrated Consistent Authority File Service for Nationwide Use. Library of Congress Information Bulletin 39(July 11, 1980) p. 244–248.

(19) Svenonius, E.: Translation between Hierarchical Structures: An Exercise in Abstract Classification. In: Proc. Third International Study Conference on Classification Research, Bombay, January 1975. Bangalore, India: DRTC 1979.

Address:
Prof. E. Svenonius GSLIS, UCLA
Los Angeles, CA, 90024, USA