CHALLENGE PAPER: HOMEOSEMY -- ON THE LINGUISTICS OF INFORMATION RETRIEVAL*

Hans Karlgren
KVAL Institute for Information Science

A Need and a Tool?

[1] Linguistics is necessary for the design of future computer-based information retrieval systems.

This is a strong claim. Not surprisingly, several documentalists take offense when linguists propose it. They find that the present design of retrieval systems is not fundamentally deficient and needs polishing rather than remaking, and/or they expect major contributions to come from people within the field rather than from more or less ignorant outsiders.

Insistence on a linguistic approach is often mistaken for an argument for use of natural language. But our claim is more fundamental. We shall make [1] stronger by adding

[2] Linguistics in information science is not restricted to possible natural language processing.

In the following we shall take for granted that the reader shares our view that

[3] mechanical retrieval systems as known today are very remote from what they could become in a foreseeable future,

and that

[4] the major restriction today is not in the amount of retrievable data, the availability of the service, or the familiarity therewith among users, but in the selectivity of the search methods.

On the contrary, the development so far has increased the amounts of data accessible as compared to the pre-computer period while something rather blunt has replaced an extraction procedure which was often, thanks to qualified and dedicated librarians, highly selective and adaptable.

---

*This paper was formulated as a challenge paper to guide the discussions. Admittedly disputable statements are preceded by bracketed numbers.

Consequently,

[5] research and experimentation should be redirected from mass processing to procedure testing, from building up still more data bases to be processed by conventional methods -- the strengths and weakness of which are well-known, to the study of small amounts of materials with more complex and less known retrieval procedures.

Accordingly,

[6] building new data bases, installing new enquiry terminals, or training users in the manipulation of existing systems should not be tolerated, unless such measures can be justified on the grounds of their immediate profits without reference to possible research merits.

This view is in agreement with the principle of transferring research resources from areas of highly predictable results to such as we know less of. If a few large-scale but predictable data-base projects per year could be eliminated, the research on crucial retrieval problems could presumably be easily financed.

This is an argument for long-range planning of this kind of research. A significant improvement of performance can certainly be achieved with any technology by making users more accustomed to the characteristics of that technology so that they play the game well. Thus, if users are taught to modify their habits of asking and of writing headings or summaries, etc., existing systems will produce better results. However, knowing how little we know about search methods and realizing how short our experience of mechanical information processing is altogether,

[7] we shall normally reject every proposal for improving the efficiency of systems essentially by changing the habits of the users on this account.

It is a good question what change of habits is essential.

The reasons for [7] are, among others, that

[8] the improvement trend produced by such means cannot be extrapolated beyond a ceiling because there exist inherent limitations,

and

[9] such a palliative treatment may be as disastrous as symptom-suppressing drugs administered to a person who is seriously ill; the temporary success may mask the need for long-range research.

Instead, the users' reluctance to comply with the formats prescribed by a system should be treated as a healthy reaction and be studied as such and not taught away.

It might be tempting to specify as a target level of ambition that

[10] a mechanical search system should perform as well on large data sets as a qualified and well-informed human does on small sets,

and that

[11] the gain of mechanization should be wider coverage and faster inclusion of new material (learning), particularly important in new interdisciplinary fields which are typically those where human oracles fail.

But this level of ambition, while unattainable in some respects, is probably too low in others. Our own experience from retrieval of passages from within a single text -- say as little as 10,000 words of legal text -- reveals that, even so, humans fail to extract all relevant items. It is almost impossible to find all relevant implicit cross references to a given passage. He who has at his finger-tips all passages from the Bible, or a set of legal codes, or a normally ill-structured computer reference manual may pass as an uncommonly learned man:

[12] document (passage) retrieval may be non-trivial even for humans when the data set is very small.

Now, granted that there is a need for some fundamental new insights in information retrieval, it is still not obvious that the linguists are those who could supply the missing tool. If documentalists accept good human performance as a challenge level there may be some aprioristic reasons for expecting linguists to have something to say; they might be expected to know a little about how humans do it. But we need more evidence than that. The linguists, clearly, carry the burden of the proof for statement [1].

Let us, already at this stage, eliminate a possible compromise stating that the kind of "linguistics" necessary for these tasks is a general science about Language, including such study of formal languages as mathematical information theory or formal logic. We do not mean merely that information science and linguistics unite in the most abstract spheres. Our issue is not whether this over-all study should be labelled linguistics or semiotics or informatics. By linguistics we mean the kind of knowledge peculiar to linguists, i.e., knowledge about certain properties of natural languages.

The Retrieval Problem

One can view document and other retrieval systems as a kind of question answering devices*.

The request can be understood as "Do you have something like XXX?" and the elicited offer may be understood as "yes, I have YYY", or "In a way. I do have YYY." We shall loosely say that the answer is identical to the question if XXX=YYY.

---

*It is true that the "request" for literature of a given kind may also be understood as a command, but that interpretation does not preclude the question status of a search question, since all questions can be described as having the deeper structure of a command to supply information.

Hans Karlgren

In an effort to distinguish between document retrieval, information retrieval in general, and other kinds of question-answering systems, we have found it useful to consider the following three kinds of question-answering systems, assuming the system at any one point in time to be deterministic in the sense that it produces only and all the same answers to any one given question.

We have:

order i: systems with a finite set of questions

order ii: systems with a finite set of answers

order iii: systems with an infinite set of answers.

Order i. This group contains systems like those for airplane booking or spare parts inventories. The possible questions are many, the usefulness of the system enormous, and the problems of design and implementation may be formidable in several aspects, but one thing about them is trivial: the relation between question and answers. The questions are all foreseen. The answers could in principle be assigned *a priori* to the questions. All refinements over mere listing could be summarized as storage technique.*

Order ii. This is the typical document retrieval system. Whatever the question, the answer is a list of document references. The set of possible answers is the set of all subsets of the set of document references. The possible answers are many but in principle known prior to the questions.

Order iii. In the general question-answering system, the answers are derived on the basis of analysis of given input statements. The system can do more than reproduce statements that have been given to it. These are the systems which are expected to say whether a given substance will resist a certain load, whether a transaction is compatible with a given contract and a given set of legal rules, etc., etc. At this point, we are deeply into artificial intelligence and information processing in general.

Now, the distinction between finite-infinite may be impressive in definitions, but systems designers derive little comfort from the finiteness of very large numbers. Very large sets are often better treated as though they were infinite. To avoid in this field a repetition of the long fruitless discussions in linguistics about the finiteness of the set of sentences, our tentative definitions above should possibly be amended so that the crucial fact is whether or not the designer can make use of the finiteness.

Another consideration which blurs the nice distinctions above is the need for interactive operation as soon as the relation between question and answer becomes at all complex. Thus, a document retrieval system aiming at selecting a subset of a given finite document file may need to enter into a dialogue with the questioner. In addition to straightforward answers such as "No, we do not have that" or "Yes, we have the following suggestions..." there are other adequate responses. In fact, we will very soon find almost the same wide range of responses to a question as in other studies of the semantics and pragmatics of questions. The system may, in one disguise or another, produce replies such as "The question is unintelligible", "Yes, but that would be at least 153 000 items; do you really

---

*Text compacting is a fascinating field in itself and does have a bearing on linguistics. But here we are rather in the domain common to linguists and others who study code design.

want them listed?", "What do you mean by XXX?" or "Do you understand 'XXX' as 'YYY without ZZZZ'?", "Your question cannot be answered as it stands but do you object if we replace it by QQQ?", etc., etc. This meta-dialog may contain a potentially infinite number of systems responses, even if the number of ultimate answers is finite.

Still hoping that our distinction is of some value we shall concentrate on order ii and we shall restrain the term information retrieval to this case of information found and lost.

In document retrieval, the crucial problem -- note the singular form -- is to match the question/request with the description stored about the document. This problem does not essentially change its character if the description is of one kind or another; the full text of the document is just a special case of a description.

If the items to be retrieved are not documents but, say, persons, patents, precedence cases, chemical substances, or processing methods, the retrieval must nevertheless operate on descriptions defining what can be offered.

[13] Thus, in any retrieval system for every retrievable item there must be an internally assigned description, an offer.

[14] The general retrieval problem is to match requests with offers.

[15] The definition of a good match is far from trivial. We need less prejudiced criteria for a good match.

[16] In particular, the definition of matches and the evaluation of goodness of fit must be independent of proposed search algorithms.

The actual design of retrieval systems requires much more than a good solution to problem [13], but that is the problem which distinguishes retrieval design from other difficult systems engineering tasks.

Much has been written about the form and manipulability of descriptions. Perhaps too little attention has been given to the contents of the description. The adequacy of the description and the agreement between request and offer are different but related matters. The former concerns the relation between the object to be retrieved and its description, the latter the relation of that description to a request.

These two relations have been particularly confused when the object to be retrieved is itself a text*, as is the case in document retrieval, which therefore is a very special case. And it is clearly one for linguists to handle whether or not they are in retrieval systems design. It is necessary to have philological knowledge about the relation heading/body of text, etc. And there are essential linguistic problems in the definition of the topic of a text; this has to do with theme/rheme relations, concepts which are defined on the sentence level but not yet on 'text' level.

---

*Obviously, a text can be described like any object by extraneous properties; one would be glad to have in a scientific system such features as "original," "elementary," "echo-paper", etc. The case which is often tacitly assumed to be the only relevant one is when only the contents of the text itself are described.

Therefore,

[17] We recommend a study of the relation between document text and document description to be made independently of algorithms for deriving a description from a text. Just as it may be fruitful to investigate what constitutes a good match without considering how the pair is found, it may be worth-while to investigate what is a good description for a given purpose without considering how it is obtained.

The book, *Linguistics and Information Science*, by Sparck Jones and Kay (1973) seems to lead up to the conclusion that systems of order ii could better be handled as special cases of systems of order iii. The restriction to prefabricated answers does not make the task essentially easier. Conversely, one could maintain that even much more powerful question-answering systems could and should be designed as systems with retrieval components. Even a specially generated answer, individually phrased for the client, may be a simple function of partial answers retrieved from a set of input items (elementary statements or "facts").

Concluding,

[18] an unprejudiced study of questions and answers is crucial,

and, in particular,

[19] the role of presuppositions must be clarified.

On the one hand, it is clear that an information retrieval attempt may have as one of its major and intended results a reformulation of the question, based on the wider knowledge contained in the system (and not only in the documents themselves, if it is a document retrieval system). There must be a means for the answering system to reject presuppositions in a question without rejecting the question altogether.

On the other hand, a questioner must be able to use a system even though he does not accept all the presuppositions built into the descriptions (which may have been phrased a long time in advance). It must be possible to come around moderate shifts in perspective, if we want the system to age slowly. This problem area deserves study in the light of the modern study of presuppositions.

A Non-Linguistic Approach

[20] Irrespective of their various algorithms, it seems that most techniques practised today are based on the following conception.

Request and offer are both phrased or rephrased in a retrieval language in which the matching (the "manipulation") is performed. (The "request language" and the "indexing language" are here seen as subsets of one operation language, just as questions and answers are subsets of one language; we disregard differences of format).

Since the original request is not always given in the retrieval language it must be translated into that language prior to searching and matching. Similarly, the original (document or other item) specification may have to be transformed into an offer expressed in the retrieval language by a translation procedure often called indexing.

We then have the schema

R->R':	the given request	is translated into an effective	request	susceptible to the manipulation needed for matching
O->O':	the given offer	is translated into an effective	offer	susceptible to the manipulation needed for matching

Subsequent manipulations operate on the R' and the O' and lead to assignments of O's to R's (or vice versa, if you prefer).

The matching of R's and O's are based on one or more of the following:

i. Identity

ii. Partial identity

iii. Some logical calculus (typically: Boolean algebra).

Even when the computations are fairly complex, consisting, say, of establishing chains of implications, they break down the comparison, in a few steps, to an assessment of identity or non-identity of primitives. These matching procedures in themselves are not regarded as linguistic procedures.

Systems of this design often do include substantial linguistic components, such as

i. Parsers or other tools for input analysis (automatic reformulation of requests and/or "automatic indexing")

ii. Linguistics-inspired design of the retrieval language.

The latter may include such "natural" features as

ii.i word order as an expression of semantic distinctions

ii.ii modifier/modified relations

ii.iii links and rolls, "cases", and other relations analogous to the semantic relations in natural language.

But the crucial problem of matching, the retrieval problem, is not treated as a linguistic problem. The underlying assumption is that once the request and offer have been transferred to the exact form which the retrieval language stipulates, they can for retrieval purposes be considered as unambiguous. The rest is a *jeu de rencontres.*

[21] With this attitude towards retrieval, the use of linguistics is optional.

The designer may or he may not permit requests and document descriptions to be written in some more or less natural language. But, that is the idea, he may also impose severe restrictions on the questioner and thereby eliminate as much as he choses of the linguistic aspects and complications. Similarly, the designer may make the retrieval language more or less natural, but, and that is the implicit assumption, he is also free to define this internal logical representation independently.

The need for identity on some level -- the R and O as wholes, parts of R and O, or primitives contained in R and O -- requires a control of the retrieval language. This control is of a destructive kind and might in travesty of Reader's Digest's well-known slogan be summarized under the exhortation Decrease Your Vocabulary (and Your Syntax).

The problem of variation in the expressions of R and of O when both mean the same or almost the same thing is met by trying to eliminate that variation. There are two kinds of such linguistic reduction:

i. Elimination of variation of expression where no difference in meaning exists: standardization of usage. Standardization is sufficient only if the system is of order i ("trivial" question-answering).

ii. Elimination of the variation of expression when the difference in meaning is small. The control then means a compulsion to use the closest expression from a permitted set (sometimes, called a thesaurus although its prime property is not to be rich but to be poor).

The translation from R to R' and from O to O' then requires an approximation procedure: instead of saying what one means, one says something one does not exactly mean in the hope that the other person who is also not saying exactly what he means will have said exactly the same.

[22] One could summarize the use of vocabulary control and other language control in order ii systems as an attempt to increase the probability of rencontre by reducing precision.

Precision is here taken in its technical meaning of degree of specification. The result will be, and equally so whether the restrictions apply to R and O or only to R' and O', that certain information will never be used and that the decision never to use it has to be made prior to searching. Whenever such decisions prove to have been premature, the selectivity of the system will be impaired.

We conclude:

[23] Language control which implies reduced precision of input data is an adequate means of eliminating chronically irrelevant information but is no general solution to the matching problem,

because, and this is a major point,

[24] in the general order ii case, it is impossible to know what information to disregard in a request or an offer until one has seen the partner of the match.

We need to make abstractions but we never know *a priori* which is the best direction to abstract into. A restricted language, to the extent it is restricted, forces us to make *a priori* decisions on what to discard. Thus, a description (A, B, C) in a simple system with a finite set of descriptors leaves undecided whether this item is a special case of AB or of AC or of BC; any one of the descriptors may, in a two-out-of-three match, be disregarded. But the system did impose on the indexer the choice of exactly A in place of whatever near-A was originally given.

A Linguistic Approach

The Dream of the Ideal Language. The approach we called non-linguistic was characterized by an effort to replace natural language by a universal exact and unambiguous representation on which a calculus could be defined. Linguists smile sadly at the new proposals for exact logical representations of the meaning of natural language texts. The dream of the ideal language spurred many ambitious attempts over the centuries, since the 17th century, if not earlier. All these great men with their fantastic systems failed, certainly not from lack of time, zeal, or genius. No modern systems engineer should take it as a personal distrust when his linguist friends tell him to give up as a bad job his design for an exact over-all representation, be it a general-purpose classification of all concepts or something else. The linguists react to proposed ideal languages more or less like physicists do when presented with another proposal for a perpetuum mobile; it is not that they would not like to have one. But there is overwhelming empirical and theoretical evidence that a rigid but yet inclusive language will never be designed:

[25] A universal linguistic perpetuum immobile is not possible.

Reasonably, then,

[26] The design of exact logical representation of knowledge, except for narrowly restricted highly specialized domains, should be encouraged no more than should perpetuum mobile construction.

Hopefully, we need not take any stand on the evasive philosophical issue of whether

[27] The meaning of an utterance in natural language can in principle be specified in terms of a finite set of semantic primitives and well-defined functions thereof.

Personally I think that [27] is exceedingly implausible, but even those linguists who do support it will presumably agree that the semantic representation postulated there is something much more complex and explicit than any representation which could be considered for retrieval purposes.

Exact Meaning Representation Difficult. Special purpose codes can, of course, be invented and have been invented for particular applications. Thus a retrieval system for chemical substances may work satisfactorily if the substances are specified with some chemical formula. But as soon as the retrieval questions expected are permitted to refer to unconventional procedures for chemical procedures or non-listed families of substances, we risk exceeding the scope of such an exact special-purpose language.

Even assuming that a sufficiently inclusive exact representation were found for a given purpose, there are other obstacles to the nonlinguistic approach:

[28] It is inconvenient for humans to write and read in a formal language.

For evidence of this statement, it should suffice to refer the reader to his own bitter experience. Even moderately complex formal systems create enormous amounts of brainpain -- and errors, and that among formally trained persons, too. Even, Boolean expressions get out of hand when the levels rise beyond three or four. And professional programmers are haunted by formal errors in programs.

This is not a plea for using natural language under all circumstances. Artificial languages could and should be improved. To me, it seems evident that the many millenia of experience built into the structure of natural languages should then be resorted to. The fatal point is that if a language, artificial or not, has enough "natural" features to be attractive to human users, it is likely to become inexact and ambiguous.

Thus, one major convenience feature is what might be called redundancy adaptation: a good margin where mistakes are likely to appear and reduced expressions elsewhere. This human-oriented feature necessarily makes the texts at least locally ambiguous.

Similarly, the semantic flexibility, which is probably necessary, is liable to produce vagueness; we shall come back to this point.

A less obvious obstacle to inducing humans to produce formalized output is that they often fail even when their performance is formally correct. Humans users tend to introduce unintentional "natural" features. Stretching the meaning of exact to something like "having the well-defined meaning which can be derived from the specification of the language, and nothing but that meaning", we could even put it

[29] An unambiguous and exact man-made text does not exist.

What I am trying to say is that one may well be cheating oneself into believing that R and O were more formally specified than then really are. The writer and any human reader will still read into the text information which, according to the definitions, of the language, is not there, and which will be ignored by the system.*

---

*If we look at all such permutations of the statements of a Fortran program as are equivalent tc the computer, only very few "make sense". On the other hand, quite a few moderately wrong programs do make sense to a human reader, who can correct them without real effort. Similarly, a mathematical proof, say, of Pythagoras' theorem, might still be a valid proof, after shuffling some of the lines, but no reader will be able to see the point.

Man-made texts seem to have a text-structure, presuppositional restrictions etc., whatever the *a priori* norm says about those.

A very simple illustration. The Boolean (sub-)expression "A and not B" may well be intended as "A without B". Now the meaning of 'A without B' is quite complex, as appears from the fact that the offer of a document carrying the title

diesel engines without injectors

is an adequate response to questions such as

diesel engines with injectors

or even

injectors for diesel engines.

For whoever wrote about diesel engines "without" injectors, presumably either explained why such deprived engines are adequate or he makes the point that injectors are not, after all, indispensible. In either case he says something important on the role of injectors in diesel engines.

The point is not that a retrieval language cannot do without without, but rather that we cannot be so sure we have eliminated it just because there are no other connectors than AND, OR and NOT in the texts. *Naturam furca expurgas...*

Some convenience can be gained by using a less formal input language for R and O and translate it into an operational language. Now,

[30] The translation of R and O, if written in a language which has enough of natural features to be attractive to human users, into an exact logical language will cause substantial losses of information.

[31] These translation losses will in general be unpredictable by the user, unless the two languages are very close to each other.

We may note that exhaustive internal recoding of a given more or less informal text is a far more advanced task than the admittedly difficult (mechanical or manual) translation between natural languages, since in the latter case obscurities may (in fact: should) be transferred unresolved to the target text.

Exact Representation Unnecessary. This may all seem defeatist. A slightly encouraging observation, however, is that there is not necessarily a real need for a complete logical analysis of the requests and offer. We are interested in their mutual agreement rather than in their explication. What we want to do, after all, is not to relate them to a system of general knowledge but to relate them to each other. We need to know how well O approximates R rather than the absolute value of either.

In numerical work, we know exactly what "approximately" means. "Appr. 1.3" may be defined as "not less than 1.25 and not more and less than 1.35" or as 1.3 plus/minus an error which can be well defined, say 1 standard deviation. We need a theory for qualitative approximation, explaining exactly what it means that an offer is approximately what was asked for. In what way is silver an approximation of gold, of money, of chrome, and of photography?

We can compare with the mathematical problem of determining the common divisor of two given numbers. We achieve this if we analyse each number by itself and check the two results for common primes. A well-known algorithm, suggested by Euclid, produces the common divisor directly, by successive operations (divisions) on the two given numbers or one of them and the result of the previous operation. We need Euclidean algorithms, operating on the pairs of an R and an O and yielding (a measure of) what is common to them. We need to establish not the meaning of one given expression but the similarity of meaning between two given expressions. The fundamental concept, then, is not meaning, but similarity of meaning. We shall use the word homeosemy (from Greek homojos, almost the same) for the similarity of meaning between two expressions.

Exact Representation Insufficient. It is often taken for granted that given exact representations, R' and O', the agreement between these two will be trivial to define (if not to find; long inference chains may present formidable computational problems). Basically, it is assumed that similarity can always be reduced to partial similarity (or to very simple logical relations; cf. supra).

We find no real support for such an assumption.

[32] If we want to define a topology over the set of expressions rather than to explicate each of them, we need not assume that the homeosemy of any two expressions or expression components can be further reduced. Rather, our analysis should be built up from the concept of distance or association between primitives ("elementary homeosemy").

Homeosemy, then, becomes a quantitative concept as fundamental as inference or set inclusion. It is in this perspective the work on associative word association should be seen. Consider our attempt at a formal mathematical formulation of association with 'damped transitivity' (Brodda and Karlgren, 1969).

I hesitated whether I should put as sub-heading for this section

[33] Exact representation is not desirable.

What I meant was that vagueness may be the price that has to be paid in order to achieve the kind of gliding from one concept to another which is necessary for non-trivial retrieval.

We note that in natural languages -- and their design is successful in this respect -- communication normally proceeds without explicit definition of terms. Not only do different persons attach slightly different meanings to the same terms but no person has ever even to himself delimited an exact or definable meaning of terms, except possibly for some few of them.

[34] In normal human communication, introduction of an explicit definition for natural language terms is a symptom of malfunction.

One may ask oneself whether natural language succeeds not in spite of but thanks to the absence of rigid definition of meaning. The flexibility of natural language semantics appears also from the observation that

[35] definitions of terms age much faster than the terms themselves.

The Wilkins' 17th century classification of elementary English terms in a systematic conceptual classification, proceeding from divine, human, animal and so forth, sounds very ancient. But the words themselves remain with approximately the same meaning and actual texts from the same period can be reasonably well understood today by those who know Modern usage.

Although we must admit that we find natural language more impressive the more we see of how it works, all this was not mentioned as an argument against artificial languages but as a reason to build into any retrieval language some of this flexibility. For this purpose,

[36] meaning differences between terms and meaning shifts should be studied, particularly the meaning shift between question and answer and the shift due to introduction of new terms, in undefined vocabularies.

Here, we can draw on rich funds of philological knowledge.

Effects of Linguistics

So Far. The effects so far have been surprisingly meager, as is made evident by Sparck Jones and Kay and by later publications. The linguistic designs which have been tested have demonstrated little effect over straightforward non-linguistic methods. In some cases the linguistic ingredient seems to have had a negative effect, even in matters of analysis of natural language input.

In principle,

[37] immediate practical tests of economy or over-all selectivity are not adequate for evaluating new methodology.

Otherwise a polished primitive system will almost always win over innovations which are less ripe for production. Practical mileage economy tests may be adequate for a Ford and a Volvo but are uninteresting in a comparison of a Volvo and a prototype electric car.

Nevertheless, when some documentalists maintain that mere recording of (truncated) terms without respect to word order or any other syntactical information performs better than linguistic analysis -- and when this happens to be true, with some qualifications -- it is a challenge.

One reason for poor success is that a kind of linguistic analysis which long ago (in some cases: centuries ago) had been given up within linguistics itself survives or is reborn in documentalist applications. Thus, models equivalent to naive dependency or phrase-structure grammar have been allowed to represent linguistics as a science.

[38] These grammars which are nevertheless in some way elementary and fundamental, cannot be used even as a first approximation to grammatical analysis. They may be good as grammar components, but a component cannot replace or approximate the whole.

Thus, an innocent analysis of 'Electronical pedagogical equipment in nursery schools' and of 'Nursery schools with pedagogical equipment' will over-emphasize the differences. It will find that different things are referred to -- gadgets and schools -- and different things are stated about them: where they are kept and how they are equipped. That kind of analysis will fail to see that two differences cancel out. The grammatical filter will remove too much in such cases but yet not disclose the similarity of such pairs as 'Finland's export to Sweden' and 'Sweden's import from Finland'.

Today. Linguistics could immediately be helpful

i. by dissuading documentalists from spending resources on vain attempts, such as

i.i creating an ideal language,

i.ii attacking the general retrieval problem by means of language reduction,

i.iii trying to make their retrieval language more stable (by means of term definitions or otherwise) instead of more flexible;

ii. by supplying

ii.i professional tools for parsing and similar tasks,

ii.ii means for synonymy manipulation,

ii.iii quantitative association methods for study of associative structures,

ii.iv *ad hoc* methods for grammatical filtering.

Since full-fledged analysis is probably not practical, algorithms must be designed on the basis of the deeper insights about, say, transformational relationships. Computational linguistics has lost its innocence and must draw some conclusions therefrom. One conclusion is that since an expression in natural language can under various assumptions yield so many reasonable interpretations, the procedure must take the uncertainty of any analysis into account.

In the Future. Linguistics must be involved in the retrieval problems as such. These are linguistic by nature; there is no choice whether or not to treat them in a linguistic manner.

Effects on Linguistics

More focus will be placed on

i. exact study of inexact expressions

ii. study of shifts of meaning

iii. study of question-answering

iv. semantic topology.

So far, emphasis has been placed on the binary distinction between the same and not the same meaning. Of old, linguists have been keen on finding distinctions which were otherwise overlooked. Lately, linguists have established equivalence classes of expressions which have exactly the same meaning, (feeling very unhappy, some of them, when these 'variants' turn out to differ after all, at least in theme/theme relations). Systematic study must be made of the agreement between such expressions as cannot be treated as semantically equivalent.

References

Brodda, B., and Karlgren, K. "Synonyms and Synonyms of Synonyms". *SMIL*, 1969, 5, 3-17.

Sparck Jones, K., and Kay, M. *Linguistics and Information Science.* New York, Academic Press, 1973.