F I R L
(Facetted Information Retrieval for Linguistics)

A DOCUMENTATION INDEXING LANGUAGE & CONVERSATIONAL QUERY SYSTEM
(FIRL -- LANGAGE DOCUMENTAIRE ET SYSTEME CONVERSATIONNEL DE QUESTION-REPONSE)

Brian Harris
Linguistics Documentation Centre, Universtiy of Ottawa
&
Eric Sinas
Ottawa-Carleton Regional Government

ABSTRACT/RESUME

(1) REPORTS IN A BRIEF MODEL FOR document indexing IN natural-language terminology BASED ON faceted classification FOR THE PURPOSE OF evaluating queries in a truly conversational 'prompting' retrieval system, EXTENDING the dimensions of faceted indexing;
(2) REPORTS A MODEL OF synonymy and hyponomy USING faceting and 'rings' FOR THE PURPOSE OF equating synonyms and translations and disambiguating polysemes, EXTENDING the ring model MADE BY the DEACON Project;
(3) REPORTS A COMPUTER PROGRAM FOR implementing the models.

(FAPS abstract)

(1) EXPOSE BRIEVEMENT UN MODELE d'analyse documentaire FONDEE SUR LA terminologie usuelle du domaine d'application, AUX FINS DE l'évaluation des questions par un véritable système conversationnel de dépistage muni d'un "souffleur", AINSI PORTANT PLUS LOIN les dimensions de l'indexage à facettes.
(2) EXPOSE UN MODELE DE synonymie et d'hyponymie A L'AIDE DE facettes et d' "anneaux" POUR établir l'équivalence entre les synonymes et les traductions et POUR distinguer entre les polysèmes, AINSI PORTANT PLUS LOIN le modèle annulaire TEL QUE RAPPORTE PAR le laboratoire DEACON.
(3) RAPPORTE EN RESUME un programme d'ordinateur POUR réaliser ces modèles.

(Résumé FAPS)

ACKNOWLEDGEMENTS

The research reported in this paper has been made possible by a grant from The Canada Council. The text itself has been set up and printed on the IBM ATS/OS system at the University of Ottawa.

1 THE LANGUAGE

1.1 Query Evaluation & Prompting

A frequent problem encountered with information-retrieval systems is that the enquirer does not frame his query precisely enough, with the result that the output contains too much noise. In terms of a document-retrieval system, this means he receives too may documents that are low on the scale of relevance. If the system were able to evaluate the query before conducting the search, if furthermore it were equipped to then ask its own questions of the enquirer in order to elicit more precision in the query, this would consititute an advance in the artificial intelligence of such systems. Adapting Turing's well known criterion for judging whether a machine can think (Turing, 1950), we might say that dialogue with such a system should convey to the enquirer the illusion that he is conversing with a human librarian.

To evaluate the query, the system has to know not only how much information it needs but also what kind --- the latter in order to ask questions back. The purpose of asking questions back is to extract a formulation of the query which is more precise or better suited to the indexing language. This is one of the ways in which FIRL has been designed to 'prompt' its users. If in addition we allow the enquirer to express himself in a comfortable subset of natural language, the system must also be capable of translating the query into the retrieval language; conversely the retrieval language must be such that it can be translated into. It may also happen that the query is not imprecise in the mind of the enquirer, but that ambiguities have crept into his linguistic expression of it. So the system must be able to detect ambiguities and ask questions that prompt the enquirer to resolve them.

1.2 Faceted Data Structure

The data structure we use is a type of 'faceted indexing'. It differs in several ways from the classic model of Ranganathan and his school on the one hand (Ranganathan, 1967), and from freer faceting such as PRECIS uses on the other

(Austin & Butcher, 1969). Syntactically there is a fixed number of facets (at present 17) and a fixed order. Any number of descriptors may be placed in a facet; but the system requires the enquirer to enter at least one descriptor in each, even if it is only "n/a" like on a tax form. That is how it decides whether it has enough information to proceed.

If the query is inadequate, the system poses prompting questions that are based in the first instance on the semantics of the facets. These semantics are worked out by linguistic study of the the scientific terminology of the discipline, which in the present experiment is limited to linguistics itself. For example one finds many compound terms such as 'computational linguistics' or 'historical phonology'. In these the head term refers to some kind of analysis that the linguist performs; the qualifier adds the general approach or methodology. There we have the semantics of two of the facets. If an enquirer asks only for PHONOLOGY, the system can now ask an appropriate question such as, "What general approach or methodology in PHONOLOGY interests you? If none in particular, answer: n/a." For other facets we have used a semantics which assumes a certain epistemology of the discipline. An explanatory list of the facets currently in use is given as Appendix 1, but there is not space in this paper to give a justification of them (Hartmann & Stork, 1972).

If the enquirer is stuck trying to answer a particular prompting question, we propose to offer a 'second level' of prompting. He will be able to call a list of terms that the system already knows to be appropriate to the facet in question, and give a 'multiple choice' response. (This kind of prompting is similar to the HELP command in the IBM STAIRS system.)

It follows that the documents in the system's data bank are indexed in the same faceted language.

1.3 Synonym Pings

The system recognizes two kinds of ambiguity. One is where a term could be placed appropriately under more that one facet. For instance, does TRANSLATION mean 'translated texts' (Facet 31: Linguistic objects), or the 'act of translating' (Facet 21: Action by or affecting the speaker)? In such cases the prompting question will be similar to what has been described.

The other kind is where more than one meaning of a homograph would be appropriate under the same facet. For example, does an enquirer who enters FORMAL mean it in the

Bloomfieldian sense or in the logico-mathematical usage (Lyons, 1968)? In either case it would fit in Facet 13: School or model. Here disambiguation depends on the dictionary structure in the system.

The dictionary not only has to cope with ambiguities, but also with synonyms. Given that our inventory of descriptors is compiled by 'literary warrant' (Foskett, 1972) and likewise that the enquirer can use the terminology of natural languages (the system is bilingual), a wide variety of synonyms and translations can be expected. The internal structure of each dictionary entry is basically a ring, so that any one term in the ring points either directly or transitively to all the other terms in it. A required element in each ring is the number(s) of the facet(s) in which it can be used. In this system, therefore, ambiguity can be defined operationally as

(i) a term belonging to a ring with more than one facet number

or (ii) a term appearing as an element in more than one ring

or both.

Assuming two rings

(1) Facet i --- a1 --- b --- c ...Facet i

,a1=a2 & b≠d

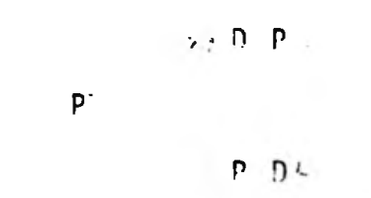
(2) Facet i --- a2 --- d --- e ...Facet i

and a query containing a, the system can now ask a prompting question: "Do you mean a in the sense of b or rather in the sense of d?" If the enquirer answers, "b", the system will obviously use a1, b and c in the search, and not a2, d, nor e. (The answer, "Both b and d" is also valid.)

For the time being construction of the synonym-cum-translation dictionary is entirely manual, as is the indexing of the documents. Automation of these operations may come in a later stage of the project.

Synonym rings were used in DEACON (Thompson, 1964). A ring represents relationships between its elements which are symmetrical and transitive, because whatever element one starts at and whichever way one goes round the ring one always reaches every other element. Such power --- a ring is more 'powerful' than a list, for example --- may sometimes be more than is desired. For instance one may want to orient the relationship in the direction general to specific, but disallow the inverse. If a query contains PRONOUN one may wish it to cover a more precise partial synonym such as DISJUNCTIVE PRONOUN and translations of the latter, eg. PRONOM DISJONCTIF; but may judge that to go in the opposite direction DP/PD to P would introduce too much noise. We have therefore extended the ring

model by the addition of centripetal tangents. The case just cited can be represented:



We refer in our own jargon to the tangential relationship, that of P in the diagram, as 'one-way' synonymy; whereas DP and PD are 'two-way' synonyms of each other.

1.4 Implications

An economy provided by the semantics is the use of implications of the general form:

Term a in Facet X implies Term b in Facet Y, $y \neq x$.
Thereupon,
for all Terms i where a implies i in a Synonym Ring in Facet X
& for all Terms j where b implies j in a Synonym Ring in Facet Y,
i implies j

By these means, if the query contains PHONETICS in Facet 11 (Linguistic analysis), the system can assume PRONUNCIATION and ARTICULATION in Facet 21 (Speakers' performance) without any prompting.

Another type of implication which we use is the insertion of certain descriptors automatically unless the analyst expressly countermands them. Thus, in the spirit of modern linguistics, SYNCHRONIC and DESCRIPTIVE are inserted automatically in Facet 12: Mode. In computer listings of the document bank (see Appendix 2) these 'implied' descriptors are represented by the symbol '='.

1.5 Operators on Descriptors

As already mentioned, several descriptors may be attached to a document under each facet. This straight away adds a dimension to faceted indexing, which has tended to be unilinear. Instead of an array

F1(d1), F2(d2)...FN(dn)

We now have

F1(d1, d2,...dn)
F2(d1, d2,...dn)
:
FN(d1, d2,...dn)

The effect can be seen in the sample entries printed as Appendix 2. It takes a computer to manipulate such an array.

Our experience so far at indexing leads us to think that this may still not be enough for satisfactory document description. It does not allow the analyst to make the distinction between documents that directly describe a phenomenon or describe a method, and those that describe or criticise other people's work on the same subject. Looking at the literature this way, we might say that there are two levels: description and metadescription. Chomsky's own analysis of language is description; Lyons' book on Chomsky is metadescription --- as indeed is Chomsky's own work on the 'Cartesian' grammarians. Of course a single document often contains a mixture of both levels; so that the analyst needs a notation which can be applied to individual descriptors. The notation we use for this purpose is square parentheses around the 'metadescriptors'. The descriptor TRANSFORMATIONAL in Facet 13 means therefore that the author of the document uses transformational models; [TRANSFORMATIONAL] means that he discusses their use by himself or by others.

Further refinements we provide are the operator '+' to put before a metadescriptor if the discussion is favorable, and conversely a '-' sign if the criticism is adverse. Although these operators cannot but add to the descriptive power of the indexing language, we are still in the dark as to whether they will add to users' satisfaction with the retrievals.

Another kind of distinction concerns the degree to which a descriptor is applicable. Indeed this is a perpetual headache for indexers: they must constantly decide whether a mention of a topic in a text is a significant mention worth indexing. A device frequently used in books is to print important references in bold-face type and less important ones in light face. Our analysts mark the distinction by placing the descriptor in round parentheses if they are in doubt about its importance for the document being indexed. We dub such a descriptor a 'marginal'.

These operators can be compounded (on [, on + or on - . Thus ([+TRANSFORMATIONAL]) means that the document discusses transformational grammar favourably but only briefly and in passing.

1.6 Distributed Relatives

Distributed relatives are concepts which are closely related but are artificially scattered by the indexing mechanism (Foskett, 1972), in our case by assignment to different facets. To overcome this scattering we provide cross-reference indices. For instance Middle English might be expressed by ENGLISH as the first descriptor in Facet 41 (Language) and 1066-1400041/1 in Facet 44 (Historical Period), where 041/1 'binds' the descriptor 1066-1400 to the first descriptor in Facet 41.

2 PROGRAMMING

2.1 System Organization

At the centre of our implementation of FIPL are 4 files for data storage and 3 packages for data manipulation, retrieval and printing. The files are stored on tape, but during execution they are loaded on a direct-access device. This part of the system is almost ready. It leaves us with the user-interface modules to be worked on during the coming Summer. Unfortunately there is not enough space in this paper for us to go into details, but some extra documentation is already available on request and a full report will be issued later.

2.2 Data input

The document bank is being entered via IBM ATS/OS, which is an on-line system for text entry and editing. The University of Ottawa's ATS contains useful commands added by the OPTICAL project. Our format can be seen in Appendix 2. We have endeavoured to make the input routine as interactive as possible, by pre-stocking a 'form' which is presented to the operator line by line for him to fill up.

2.3 Language & Hardware

The language being used is PL/1, (except for 2.2). Although rather slow for realistic retrieval times, it is, of the languages available to us, the closest high-level language to the assembler language. It must be borne in mind that this is an experimental project: the aim is to develop and test the power of the system --- speed is secondary for the moment. The operating system being used for the time being is IBM 360/65 TSS (Time Sharing Systems) on an IBM 360/65 of the National Research Council of Canada, with some preliminary batch testing

on another 360/65 at the University of Ottawa. Access is from IBM 2741 terminals.

REFERENCES

AUSTIN D, BUTCHER P. PRECIS: a rotated subject index system. London, British National Bibliography Marc Documentation Service, publication No. 3, 1969.

FOSKETT A C. The Subject Approach to Information. 2nd edn. Hamden, Conn., Linnet Books & Clive Bingley, 1972.

LYONS J. Introduction to Theoretical Linguistics. London, Cambridge UP, 1968.

RANGANATHAN S R. Prolegomena to Library Classification. 3rd edn. Asia Publishing House, 1967.

THOMPSON F B et alia. Deacon breadboard summary. PME4TMP-9. Santa Barbara, General Electric Co, 1964.

TURING A. 'Computing Machinery and Intelligence'. Mind. Vol. 59, No. 36, p. 433-460.

For explanations of the linguistics terminology used in this paper, a good reference is:

HARTMANN R R K, STORK F C. Dictionary of Language and Linguistics. London, Applied Science Publishers, 1972.

APPENDIX 1

FIPL FACETS

11. Field of linguistic analysis, ex. PHONETICS, BILINGUALISM.

12. 'Mode', ie. general approach: often expressed as an epithet of 11, ex. COMPUTATIONAL, HISTORICAL.

13. Model, school or method, ex. TRANSFORMATIONAL, AUDIO-VISUAL.

14. Interface with other disciplines, ex. SOCIOLOGY, CURRICULUM STUDIES.

21. Elements of speakers' linguistic performance, ex. ARTICULATION, SPELLING ERROR.

22. Attributes of speakers affecting but not part of their performance, ex. CHILD, APHASIC.

31. Language objects, i.e. elements of the language as analyzed by linguists, ex. PRONOUN, TRANSFORMATION.

41. Language group or specific language, ex. INDO-EUROPEAN, PIDGIN ENGLISH.

42. Register (the French 'niveau de langue'), ex. CONVERSATIONAL, LITERARY.

43. Dialect, regional or social, ex. QUEBECOIS, CANT.

44. Historical period, ex. CONTEMPORARY, 1066-1400.

51. Document type, ex. JOURNAL ARTICLE.

52. Level of intended readership, ex. POSTGRADUATE.

53. Language of publication, ex. RUSSIAN

54. Author's surname.

55. Year of publication.

56. Number of pages.

FIPL

APPENDIX 2

EXAMPLES FROM COMPUTER-STORED DOCUMENT BANK

```
$$
* INSERT IDENTIFICATION: Derric J, 'Listening to language
in the infant school', "English for Immigrants", (4)1971, 2, 17-22
$$ * INSERT DESCRIPTORS: 11: =, pediolinguistics
. 12: =, [didactic], experimental
. 13: [+informal], pattern drill
. 14: (sociology)
. 21: second language learning, interest
. 22: n/a
. 31- conversation
. 41: English
. 42: immigrant
. 43: England
. 44: =
. 51: article
. 52: postgraduate
. 53: English
. 54: Derrick
. 55: 1971
. 56: 5
. $$ * INSERT ONE-WAY SYNONYMS:  21: motivation, interest.
$$ * INSERT TWO-WAY SYNONYMS:
11: pediolinguistics, paediolinguistics.


$$
* INSERT IDENTIFICATION: Lewis D, 'Problems of bilingualism
in Papua and New Guinea', "Kivung", 4(1971), 1, 21-29
$$ * INSERT DESCRIPTORS: 11: =, bilingualism
. 12: =, [didactic], contrastive
. 13: [socio-cultural], [@11/2 coordinate], [@11/2 compound]
. 14: (economics), (psychology)
. 21: speaking, literacy, motivation
. 22: n/a
. 31: n/a
. 41: (French), English, vernacular, (Melanesian Pidgin),
lingua franca
. 42: home, school, (rural), (urban)
. 43: (Canada), Papua, New Guinea, (Quebec)
. 44: =
. 51: article
. 52: postgraduate
. 53: English
. 54: Lewis
. 55: 1971
. 56: 8
.
$$ * INSERT TWO-WAY SYNONYMS:
13: coordinate, coordinate bilingualism.
13: compound, compound bilingualism.
42: rural, country.  42: urban, town.
```