1. Introduction

We present here one aspect of a research project investigating the ways in which users of computer manuals and text-based information systems obtain information from these sources and build up their knowledge structures in a specialised domain (computer security). The research is primarily concerned with the early formulation of knowledge needs; the aim is to map users' needs, together with the terminology which helps to express them, onto the specialised language of computer manuals. If we come to understand this relationship, this mapping (see also Kukulska-Hulme, 1992), we are then better equipped to design retrieval tools which will make learning from texts easier and more effective.

2. Specialised texts and their readers

An age-old problem, shared by writers and readers alike, is the logistical impossibility of satisfying the information needs of all readers by the production of a single text. This explosive relationship of "one to many" is waiting to be counterbalanced by a means of transforming a ubiquitous text into an interpretation by, or for, an individual. A given person may be prepared to spend time interpreting texts, but overall the cost of the interpretation effort is running high, and access to knowledge in texts has to be made easier. In particular, training in new domains of knowledge, such as those related to new technologies, is continually on the increase, with high training costs and the scarcity of suitable manpower swaying decisions in favour of self-teaching from texts. In this scenario, specialist information which may be obtained from technical manuals increases in value, and manuals come to play a dual-reference and training-role.

As writers' skills become augmented by new computerised tools which can analyse and provide assistance with their writing, so too readers' expectations of a higher standard of documentation are growing. If a higher standard is not delivered, or if a reader has doubts about the precise meaning of parts of a text, there is still potentially the opportunity to look at a manual through a comprehensive or retrieval tool which would aid and support individual learning and understanding. The emphasis here is on the fact that a text needs a reader in order that the communication process might be complete; appropriate tools can make it possible for readers to assume conscious control of the way they "complete" the texts they read. What matters here is the recognition of the inherent imperfection of texts, and at the same time the realisation that readers are not powerless when faced with this fact. Texts may not be the most effective way of communicating specialist knowledge, but the fact is, they are still the most common way of reaching a wide and varied audience. The problem of "knowledge transfer" or the effective communication of specialist knowledge through the medium of texts depends crucially on the question of whether texts can be made to respond to their readers' needs.

Specialised knowledge is not confined to specialists; computer users are keenly aware of this. In order to satisfy the requirements of their job, which itself may be highly specialised, users have to take on board the special knowledge which will enable them to handle computers in an effective way. Often with limited training, they tackle applications as best they know how, turning to colleagues and printed or on-line sources (e.g. manuals, help facilities) when problems occur. Computer security, a domain of knowledge which cuts across computer applications, software and hardware systems, is an area of computing which many users, sometimes at their peril, ignore.

3. Terminology in retrieval

The advent of widely available computerised retrieval systems has brought into sharp focus the question of which terms are best used for information retrieval - both traditional and computerised retrieval. Even on the part of computer scientists, there is now a much firmer conviction that "information retrieval systems are fundamentally linguistic" (Blair, 1992:200), prompting a reassessment of language theories and philosophies (e.g. those of Austin, Searle and Wittgenstein) in this light. Our particular orientation brings into focus the contribution that terminological study can make to the understanding of human processing of specialised texts, making an explicit link between terminology and knowledge in the context of comprehension. Above all, we are exploring the role of terminology as a means of information retrieval, and noting in particular that domain knowledge can be acquired by repeated information retrieval, whereby a picture of the domain is built up gradually. Information retrieval is thus part of the process of "comprehending" the domain.

We are considering the computer manual as a learning tool, so that every instance of information retrieval from the manual contributes to the process of building up knowledge structures in the mind of the reader: "Every search for information is a learning process" (Alberico & Micco, 1990:17). When we examine the domain of computer security, we can see that it has two aspects: preventive and remedial. Prevention involves risk assessment and planning, and this must be done on the basis of knowledge of the domain, rather than on the basis of separate pieces of information. Remedial action resulting from security breaches means tackling the problems associated with that breach, and also implementing corrective measures to ensure future security. Any information retrieval from a manual dealing with security issues thus takes place in relation to knowledge of the domain, and provides the inputs to a process of reasoning about the domain.
It is widely recognised that a number of problems associated with computer manuals remain to be addressed. It is common knowledge that typically, manuals are resorted to when other sources are unavailable or lack the necessary knowledge, rather than being the first port of call. The difficulties in finding the right information are partly terminological in nature. In particular, there are problems at the level of entry to the text (e.g., via the index and table of contents). For example, consider these words which spring to mind when one thinks of computer security: insurance, disaster, piracy, fire protection, legal requirements, power supply, classified information, illegal sign-on. These may not figure at all in a manual on computer security - not even in the index - which suggests that a casual user might have difficulty in obtaining information from the manual. The need for a user and language-oriented approach is confirmed by Smeaton (1992) when he writes about the current limited role of Natural Language Processing techniques in information retrieval research:

"... This role does not really address issues of retrieving information for users based on the language used in queries or in texts. Fundamental issues and questions dealing with the notion of a retrieval model and document relevance will need to be integrated with what NLP techniques have to offer if really significant progress in retrieval effectiveness is to be expected."

(Smeaton, 1992:277)

4. **Control knowledge**

Researchers working on knowledge representation systems make use of the notion of *control knowledge*. This is knowledge that e.g. a rule based (production) system must have to control the "fringe" of rules: "Meta-rules are rules which control the use of domain rules" (Williams & Bainbridge, 1988:106), Control, or *strategic knowledge* (Clancey, 1983), specifies the ways in which knowledge elements are used; it also describes WHEN and WHY rules should be applied. "Control knowledge describes what to do, when and why (knowledge which can be extremely important to a novice)" states Anna Hart (Hart, 1989:59). It seems that the notion of control in relation to knowledge systems could be a very important one, given that "the most essential factor in the dynamic operation of a system is control, definable as a process for managing the relation between determinacy and indeterminacy" (de Beaugrande, 1988:9).

5. **Control knowledge and terminology**

From a terminological perspective, it is concepts, not terms, that constitute units of knowledge. As has been pointed out by Sager (1990), concepts are notoriously difficult to define; it is, however, possible to group them into four basic types: *class concepts* (e.g. entities, generally corresponding to nouns), *property concepts* (e.g. qualities, for the most part corresponding to adjectives), *relation concepts* (realised through different parts of speech), and *function concepts* (e.g. activities, corresponding to nouns and verbs) - [see Kukulska-Hulme (1989) for an account of an analogous grouping in the design of dictionaries for translators]. We can consider systems of concepts and endeavour to specify relationships between concepts, uncovering the knowledge structures which bind them together. But we cannot do the same with terms. Terms are existential in nature, that is to say they signal the existence of an entity, a relationship, an activity, a quality. Considered outside of verbal context, they can express neither facts nor rules, they can say nothing about the manipulation of knowledge. Furthermore, many concepts, particularly of the relational and functional types, are designated by words of general, not special, reference. This brings us to the crucial question: in a situation of knowledge retrieval from a written source, how can a terminological unit represent a specific knowledge need, other than one which seeks only to discover the concept (meaning) of a specific term or to confirm the existence of a concept? The procedural, functional, relational, control aspects of knowledge are not well served by terms as symbols of knowledge units outside of context. This must be borne in mind when considering the process of retrieval and knowledge transfer.

6. **Knowledge representation through verbs**

As Smeaton (1992) has pointed out,

"It has always been assumed by researchers that in language it is the noun phrases that are the content-bearing units of information. This is not true for a full representation of meaning but noun phrases are good indicators of text content and for traditional information retrieval, that is what is wanted."

(Smeaton, 1992:272)

The emphasis on "noun phrases" and thereby on class concepts in preference to other concept types may indeed be appropriate for traditional information retrieval, retrieval which presents the user with a list of references as its outcome. But if in full text-based systems, one has to keep an open mind as to which concepts will be the most effective, the most representative of a user's intention. The ultimate goal of information retrieval from a computer manual is not merely to extract information, but to gain understanding and to do something with the knowledge gained - to act in an informed way. "Comprehension is less a matter of being able to reproduce the facts in a text than of what one does or is able to do as a consequence of interacting with the structure of the text." (Smith, 1982:85). The success of information retrieval in terms of comprehension and action depends on the questions that an individual asks. This fits in well with the function-content approach to cognition favoured by Yale (Galambos et al., 1986), where the emphasis is on the content of cognitive activities - motivation, plans, goals, and outcomes -, and on function - how content will be used subsequently in tasks. In this light, we can view retrieval as a goal-oriented cognitive activity related to real world tasks.

If we recognise that retrieval is related to the need for control knowledge and to subsequent action, this may be sufficient grounds to cast doubt on the tendency in information retrieval to discount verb forms as descriptors in favour of nouns. The advice proffered by information scientists for the selection of thesaurus terms has for a long time been: "Terms should be in noun form, and verbs should be avoided." (Atchison & Gilchrist, 1972:14). Of course, it is necessary to match retrieval terms to terms in a body of specialised texts, which are seen to be predominantly nominal. But if we shift the focus from text to user, we are faced with the possibility that needs might be centred on verbs if they are concerned with action. The same authors state that "... terms arising from questions likely
to be put to the system are as important as those taken from the literature... The questions should be collected from users or from records of questions already encountered" (Atchison & Gilchrist, 1972:69). In any analysis of users' queries and their use of terminology, we have to bear in mind the knowledge needs which underlie the query.

7. **Empirical research**

A survey of computer users was conducted and has yielded some interesting data, including examples of specific user queries in the domain of computer security on the IBM AS/400 computer system. The survey also collected information about users' previous experience with computers, their prior knowledge of the domain, and their use of various information sources.

The data concerned with user queries, which is now being analyzed, consists of direct questions, for example: What security levels are available on the system? Can a user have control over his/her own profile? How many times should a user be allowed to try and sign on before the device becomes in active? Can a user sign on to the system remotely? What are the rules on password reuse? Is there an audit trail to show who has made changes to user profiles? Can the password file be decrypted so that it may be read by potential hackers?... and so on.

The query data is being analyzed in several ways, namely: conceptual/rhetorical, lexical/terminological and collocational. The first of these tries to identify clusters of questions which express similar knowledge needs. Rhetorical words such as 'how', 'what', or 'when' are a starting point for this classification process, and the result is a map of knowledge needs relating to actions and events, and questioning their timing, manner, location, cause, advisability, etc. What emerges is a picture of the respondent's perspective on the computer system, which is that of wanting to exercise control over its various components - the system itself, its users, and external influences which impact upon the system and its security arrangements.

8. **Presentation of analytical classification**

The analysis which is presented in more detail below is the lexical/terminological one. It begins with a selective frequency wordlist and moves on to a classification by four main word types: noun forms, verb forms, adjectives and adverbs. Several observations can be made about the words used in the sample. One noticeable feature is the relative abundance and variety of verbs, even allowing for the fact that verbs take on more forms than nouns. This suggests that the needs expressed in the questions are strongly associated with **knowledge about actions or events**. The suggestion is further supported by the fact that a substantial number of the nouns in the data are derived from verbs and indicate an action or the result of an action. In addition, many of the forms occurring as nouns also occur as verbs, or have the potential to function as verbs. Most of the verbs express active control over actions or events, e.g.:

- prevent
- restrict
- secure (against)
- preserve
- change
- monitor
- protect
- ensure
- increase/decrease (access)
- conceal
- make sure
- deny (access)
- reinstate
- detect
- remove

A smaller number reflect observed or anticipated events, or accidental actions, e.g.:

- hack
- break
- raid
- breach
- read
- gain (access)
- decrypt
- access
- dial in
- contradict
- corrupt
- travel
- tender
- pass through
- spread

Some indicate possibility, advisability, obligation:

- necessitate
- be able (can, can't, could ...)
- need
- have to (must ...)
- require
- ought to (should ...)

In terms of frequency, the most prominent verb form clusters signal POSSIBILITY or ADVISABILITY, and there is a bias towards active verbs.

Within the nouns occurring in the data, we can identify the following categories of words:

**ABSTRACT NOUNS OF A GENERAL NATURE**, e.g. difference, importance, extent, part, way, method...

**NOUNS RELATING TO BUSINESS FUNCTIONS**, e.g. company, contract, clause, employment, overtime, payroll...

**NOUNS RELATING TO PEOPLE AND ORGANISATIONS**, e.g. users, staff, people, personnel, security officer...

**NOUNS RELATING TO COMPUTING AND COMPUTER APPLICATIONS**, e.g. system, data, PC, computer, machine, terminals...

**NOUNS RELATING TO SYSTEM SECURITY**, e.g. security, password, virus, security level, authority...

**NOUNS RELATING TO BUSINESS SECURITY**, e.g. disaster, health and safety, precaution, protection, right...

The set of nouns occurring in the data represents, from a semantic perspective, a fairly close-knit network. The network covers, predictably enough, computing terms of various degrees of specialisation (with 'help' at one end of the scale, 'EDI' at the other), and security terms which relate to the functioning of the computer system. It also covers business concerns in a wider sense, inasmuch as the functioning of the system affects the functioning of the business, and has particular implications for its staff and for its legal ramifications.

Of particular interest are the abstract nouns of a general nature which give us some clue as to the types of knowledge represented in the data. Two meanings dominate:
a concern about HOW actions are to be performed ('way', 'method', 'measures', 'practice', 'procedure', 'steps')

- a preoccupation with timing - WHEN ('frequency', 'time', 'hours', 'interval', 'period', 'month', 're-use').

The adjectives present in the data suggest knowledge needs which seek to establish distinctions by way of contrast and comparison: the evidence for this may be found firstly in the presence of opposite pairs, e.g. 'total', 'comprehensive' vs. 'specific', 'particular'; 'secure' vs. 'vulnerable', etc.

Secondly, there are a number of negations which imply an opposite, e.g. 'non standard' => 'standard'. Comparative and superlative adjectives are also in evidence, e.g. 'greater' (greater authority, greater access), 'better' (better methods).

Adjectives may also be classified in categories which mirror the classification of nouns (ADJECTIVES OF A GENERAL NATURE, ADJECTIVES RELATING TO BUSINESS FUNCTION, ADJECTIVES RELATING TO PEOPLE AND ORGANISATIONS ...). Other adjectives concern TIMING (including repeating or lasting features), and POSSIBILITY or NECESSITY.

Adverbs: these again reflect the types of knowledge needs which have been shown to recur in the data, especially TIMING - e.g. frequently, often, periodically ... MANNER - e.g. completely, best, remotely, inadvertently ..., COMPARISON ('than', 'very'), and RESTRICTION ('with', 'without', 'no', 'non', 'not', 'only').

9. Conclusions

This analysis has allowed us to discover the knowledge needs expressed through the choice of words made by respondents to the survey. Certain concerns surface again and again in the data, particularly the need for knowledge about:

- TIMING, MANNER, POSSIBILITY, ADVISABILITY, NECESSITY and OBLIGATION

Linked to POSSIBILITY are the concepts of CHANCE or HAZARD, and the notion of CONTROL. CONTRAST, COMPARISON and RESTRICTION have been noted as specific strategies for eliciting knowledge.

Several domains of knowledge have been identified, as it has become ever clearer that computer security in a business environment cannot be considered in isolation from its closely related - broader, and overlapping - domains: business and computing. Computer security has been shown to be particularly concerned with people. This lexical/terminological analysis confirms the findings of the conceptual/rhetorical analysis, where much the same knowledge needs were signalled.

It is hoped that the final results of these investigations in progress can be used to improve the design of search and retrieval tools and indexes, whether manual

(printed) or computerised. It is only by understanding the knowledge needs of users that we can select the terms which will provide the right 'bridge' between texts and user queries.

Bibliography


