Terminological Data Banks: a model for a British Linguistic Data Bank (LDB)

John McNaught

Centre for Computational Linguistics, UMIST

Paper presented at the Aslib Technical Translation Group conference and exhibition, London 20 November 1980

A description of a model linguistic data bank (LDB) for a British market will be given, based on the results of a continuing feasibility study. A LDB represents an economical and highly efficient way of organizing Britain's efforts in the field of terminology, both with respect to English and the many foreign languages through which contact is maintained with non-English speaking countries. The institutional and organizational structure will be outlined. Emphasis will be placed on services to be provided to various groups, and in particular to translators, and on the important role these groups will play in assuring the continued viability and relevance of the LDB, not only as users, but as contributors and advisers. Data acquisition policy and financial aspects will be considered.

A multilingual, multidisciplinary British LDB will provide translators with a valuable service, whose applications are many, whose products are varied to cater for a wide range of needs, whose terminology is continually revised and updated and whose modes of consultation are several.

THIS PAPER IS based on results obtained from a continuing feasibility study of the establishment of a terminological data bank in the United Kingdom, a study being carried out at UMIST under the auspices of the British Library.

I shall use the term Linguistic Data Bank (or LDB) in preference to Terminological Data Bank, as many of the banks we investigated in the course of this study do not restrict themselves to handling terminological data alone. Thus LDB represents a more accurate designation of the types of information systems we will be discussing.

I shall concentrate primarily on work being done in this country towards the establishment of a British LDB, but shall make reference to other LDBs abroad by way of exemplification and illustration. Indeed, I would urge you to keep in mind during this talk that, when I describe possible features of a British LDB, these features already exist in other LDBs. I am not describing services or facilities or search methods that could exist. In our proposals for a model of a British LDB, we have translated the assumedly best features of LDBs abroad to the context of a British market. Where Britain may hope to achieve a measure of innovation in LDB operation is in the use of the most up-to-date technology and software, exploiting information networks and the move towards office and home computers, etc, and in reaping the benefits of recent terminological research. There are significant advantages to be gained by being

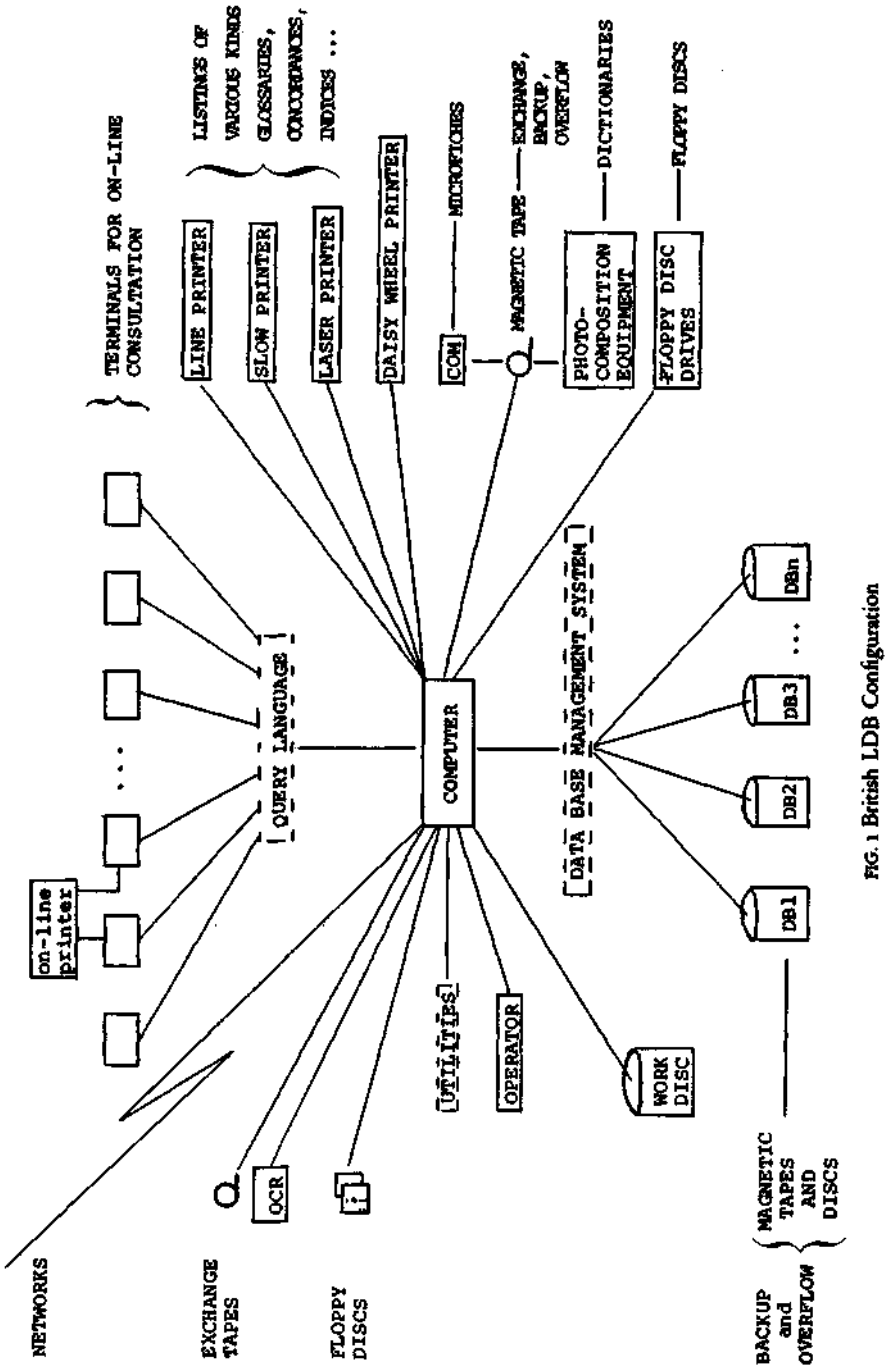

FIG. 1 British LDB Configuration

a late-comer in this field, not the least of which is to be able to study the reaction of users to existing LDBs, and so to be able to design a LDB which will suit users' needs.

There are three sections to this paper: Part I deals with the reasons behind the feasibility study; Part II is a description of the phases of the study; and Part III is a presentation of a model for a British Linguistic Data Bank.

I. Reasons

The reasons and considerations behind the feasibility study are several—I shall mention only the most important:

Special language communication. This involves the constant creation of terms to designate concepts, objects, measurements, products, etc. These designations (terms) differ from the words of general language, in that they refer more specifically than words, in that they are mainly used by specialists, in that they are often created according to established patterns and precedents, in that they are susceptible to standardization and in that they may be relatively short-lived and changed in the light of discoveries and developments.

Efficient communication. This depends on common agreement, and can only be achieved by widespread knowledge of terms (in our case) or by easy access to terminological information. The problems of efficient communication apply with even greater force across language boundaries.

Efficient special language communication. There are many different groups involved in the use and creation of terminology; all groups must have access to terminologies, both their own, and those of other disciplines.

'Information explosion'. The immense upsurge in technological innovation and the concomitant upsurge in new terminology, together with the great increase in multi-lingual communication needs, means that the work of collecting, storing, sorting and disseminating terminology cannot be carried out efficiently by dispersed methods, especially when contact must be maintained with LDBs abroad housing foreign language data.

Lack of single authoritative organization in the UK. There is no single organization in the UK able to provide authoritative guidance on English usage of specialized terminology. Note that I do not say standardized terminology: the BSI do a laudable job in this area. Specialized terminology, however, is another matter, in that both standardized and non-standardized terms are present. One is dealing with the special languages of different disciplines, with the grey areas where the terminologies of disciplines meet, with in-house usage vis-a-vis wider usage, etc. There is no national centre for terminology, no centre which has close links with other bodies concerned with the production and regularization of usage of specialized terminology. There is also a distinct lack of links with foreign LDBs—no central body capable of negotiating the exchange of data with a foreign LDB, for example.

Existence of other LDBs. In recent years, major industrial countries and international organizations have established LDBs. LDBs in multilingual form exist in (nos. of main LDBs in brackets) Canada (2), at the Commission of the European Communities, in France (1), the Federal Republic of Germany (4), the German Democratic Republic

(1), Sweden (1) and the USSR (2). In Denmark, plans are well advanced for the establishment of DANTERM. The UN plans to establish its own LDB, as does UNI, the Italian Standards Institution. In Spain, HISPANOTERM is of recent creation. Further information on these LDBs may be gained from Sager & McNaught[1]. Great Britain is the only major industrial nation without such a service facility, that is, a centre for the processing of all kinds of terminological data.

There is a substantial amount of work being done in Britain, however, related to thesauri for indexing and retrieval purposes. One of the most important contributions Britain has made in this field is towards the development of the ISONET thesaurus, which is a computerized, controlled vocabulary of some 11.5 thousand descriptors and 5.5 thousand non-descriptors used for the selection of descriptors for indexing and searching standards and technical regulations on ISONET databases. The thesaurus consists of a classified subject display and an alphabetical list (the index to the display) and, though developed at the moment only as a bilingual English-French version, is designed to be both multilingual as well as multidisciplinary. The BSI team responsible for the development of the English part of the thesaurus has helped to produce not only an excellent indexing and information retrieval tool, but also a database whose contents contain a valuable store of terminological information.

English terminology. All the foreign LDBs mentioned contain, or will contain, substantial amounts of English terminology, at least as translation equivalents, and such vocabulary may be misleading. The impact of LDBs on the usage of English terminology outside the UK will increase, and may, without British involvement, introduce usage unacceptable or even incomprehensible to this country.

There is a serious danger that the international role of English as a means of communication may be impaired if a single, national British centre for terminology does not exist. Moreover, as many languages create new terms on the basis of English, uncontrolled elaboration of English terminology in a number of different centres has far-researching consequences for effective communication in other languages and between these languages and English.

Nairobi Recommendation of UNESCO. Paragraph 12 of this document, on the legal protection of translators and translations, reads:

'12. Member states should consider organizing terminology centres which might be encouraged to undertake the following activities:

(a) communicating to translators current information concerning terminology required by them in the general course of their work;

(b) collaborating closely with terminology centres and developing the internationalization of scientific and technical terminology so as to facilitate the task of translators.'

Aslib 1978 conference on 'Translating and the Computer'. The audience of this conference expressed a strong interest in LDBs, and many of the organizations we have contacted during the course of this study were represented at this conference.

II. Phases of the Feasibility Study

On the basis of the above reasons and considerations, the project seeks to establish the following:

In phase one:

— the use made of LDBs in other countries
— the cost and financing of other LDBs
— the institutional and organizational framework of other LDBs
— the availability and quality of data for a British LDB

In phase two:

— the possible uses of a LDB in the UK
— the possible structure of a British LDB

The study itself was split into three phases:

Phase One: LDBs and data

1. The state of LDBs
1.1. Information gathering
1.1.1. Scrutiny of available documentation
1.1.2. Formulation of further enquiries to be made
1.1.3. Follow up enquiries by questionnaire or visits to selected LDBs
1.2. Report on selected LDBs:
their use, cost, financing, organization and institutional framework

Phase Two: Preliminary enquiry among potential users

1. Preliminary technical specification of a British LDB
— scope of holdings, acquisition policy
— format of holdings
— modes of operation, user facilities
— maintenance and development
2. Discussion of this model with potential contributors and users
— government departments
— relevant institutions
— industry
— translators
— information and documentation centres
— publishers
3. Evaluation of responses

Phase Three: Feasibility report on a model of a LDB

1. Modification of the technical specification
2. Organizational specification
3. Recommendations

Phase One, the material for which was gathered with the aid of a British Library Overseas Study Visit Grant, was concluded in June with a report entitled 'Survey of Five Linguistic Data Banks'[2]. Some of the main points and conclusions of this report are detailed below:

Three main types of LDB exist:

(a) those conceived primarily as translation aids, including EURODICAUTOM (CEC) and LEXIS (Bundessprachenamt)

(b) those used primarily as language planning aids, for example BTQ (Banque de Terminologie du Québec) and TERMIUM (University of Montreal)
(c) those used as aids to standardization, including NORMATERM (Association Française de Normalisation), TEAM (Siemens ag) in collaboration with DIN (Deutsches Institut für Normung), which now has its own LDB and document retrieval system (DITR) and TERMDOK (Tekniska nomenklaturcentralen), Stockholm, which collaborates very closely with SIS, the Swedish Standards Institution.

Two main methodological approaches to LDB data organization exist, exemplified by EURODICAUTOM on the one hand, which stores keywords and their contexts, in the belief that translators are best served by supplying them with terms in context, and LEXIS on the other, which records terms in isolation, preferring to work from concepts.

The facilities, services, institutional and organizational structure of these major European banks were investigated, as was the functioning of other major LDBs in Europe and elsewhere, through consultation of the literature and via correspondence.

Of great interest to us were the various systems used by LDBs to finance their operations, and to establish links with their users. Here we investigated the partnership systems set up by TEAM and TERMDOK, where partners contribute terminology in return for services, and subscriber systems such as the one operated by NORMATERM. Links with users, and methods of elaborating terminology, were studied especially in relation to TEAM, TERMDOK and DANTERM. This latter has a policy of sending terminologists into the field to develop and research terminology on the spot. TERMDOK has a smoothly-running system of committees which elaborate new terminology in conjunction with industry, etc, and has wide user links in many sectors. TEAM provides a good example of how a partnership system may operate to the benefit of all members. This particular partnership system unites many different groups and organizations, both in West Germany and in other countries, eg Philips, and the Dutch Foreign Ministry. These groups all contribute terminology to TEAM and have access to all TEAM terminology free of charge, payment only being asked for actual processing time.

In the light of the above-mentioned reasons and considerations, and given the interest manifested by many different types of user, the preliminary proposal for a British LDB is not for a LDB conceived primarily for translators, or standardization specialists, but for a LDB that will serve a wide range of users, and provide a wide range of services. This proposal is also based on the analysis of results from Phase I, where a trend was perceived among well-established banks to move towards providing a wider variety of services to a wider number of user groups: TERMDOK, for example, has recently converted to a large multi-user online system, in order to serve an ever widening range of users; EURODICAUTOM, now available on EURONET-DIANE, is now expanding to meet varied demands. TEAM system was among the first to realize the need for and benefit to be gained from serving different types of user, and the success of this system, with its many partners active in contributing terminology in many fields, and its diversified services, catering for translators, publishers, standardization specialists, information scientists and language teachers, has been a great inspiration to us.

There is a concomitant trend for proposed or newly-created LDBs to emphasize multifunctional and multidisciplinary aspects, eg DANTERM, which intends serving translators, technical writers, standardization specialists, publishers, students and teachers of Danish Schools of Economics, and other institutes of higher learning.

III. The model

We have thus proposed that a British LDB be established with the following characteristics:

— multifunctional
— multilingual
— multidisciplinary
— widely accessible

The advantages of such a LDB are:

— increased reliability and accuracy of data
— production at little cost of a great variety of up-to-date glossaries and dictionaries
— direct consultation on/off-line by organizations and individuals
— agreement can be reached and maintained between English usage of terminology at home and abroad
— a greater inflow of literature in foreign languages will be generated, which in turn will generate more demand for translations
— increased and more effective communication with foreign countries, with direct benefits for exporting, especially.

It would seem, on the face of it, that these characteristics and advantages are viewed wholly from an organizational point of view. However, every effort has been made to ensure user orientation of the LDB remains paramount. Without users taking an interest in the creation, development and running of services, an LDB will be a white elephant. Users are the life-blood of a LDB, not just in the role of end-users, but in the role of contributors and advisers. Wide user involvement will ensure that terminology is acquired, elaborated and disseminated in fields and in languages of immediate relevance to users, that services provided will be relevant to user's needs, and, with online searching and input, be as 'user friendly' as possible, and that appropriate measures may be taken to take into account origin of data, conditions of use, and copyright.

During the second phase of our study, then, we concentrated on the needs and expectations of users. We approached those people who would be likely to use a LDB (ie the staff translator as opposed to the company chairman), and supplied them with documentation on existing LDBs and preliminary specifications of a British LDB for discussion purposes. We then sent them a detailed questionnaire. Follow-up enquiries were then made, as many as time and manpower resources would permit.

Results enabled us to construct typical 'user profiles': type of work carried out, the manner in which people worked, the subject areas covered, the search and output facilities desired, etc.

Comments were also obtained on the organizational and institutional structure of a LDB, on how a LDB should be financed, on which areas of terminology should receive prime attention, on which languages should be developed, on exactly what information different users expected to obtain.

Our final report to the British Library will therefore express the wishes of potential users. LNB data acquisition policy and services should be guided by those who will use the LDB, not imposed from above.

Uses

In order to cater for numerous different user groups, uses and services of the proposed LDB should be several and varied. The main aim is to provide complete flexibility of search and output facilities. It is proposed that users should not receive exhaustive information on a term, as normally they work at any one time with subsets of term record fields. Study of user profiles reveals that certain user groups prefer to work with certain fields. Thus it is proposed that pre-specified 'packages' be offered eg term + translation equivalent + source, or term + definition, or term + translation equivalent + context, and so on. We are grateful to Mr Arthern for advice regarding such packages[3]. Also, we propose that users should be able to define their own search and output facilities from among those available, thus a request from user X would produce by default output in the format he has previously specified, information which the computer can gain through for example inspection of formats associated with user identification codes.

When working conversationally, output can be given in graduated form, a refinement of the 'package' technique. That is, one may be interested in receiving primarily term + translation equivalent. In many cases, this information may be enough for one's particular purpose. However, in case of doubt, one should be able to receive further information, simply by pressing a button, eg source + context. Which information, how much and in which order, are choices that should be left to the user. Search operations so far defined are:

(1) single term search ie a defined sequence of characters (this could be a Uniterm or a multiterm)
(2) arbitrary string search eg one may wish to output all terms beginning with, or containing, a certain sequence, for example, 'ethyl' or 'inter'.
(3) abbreviation
(4) list of terms, where information common to all terms in the list is required, eg one may wish to see whether a list of terms have the same source, or perhaps the same synonyms.

Numbers 1 and 3 involve searches of specific fields, whereas numbers 2 and 4 involve general field searches. These may be undertaken in either online or offline mode.

Online conversational mode should also allow:

(1) paging in the alphabetic order of the source or target language (paging is equivalent to browsing through the data base)
(2) paging in the systematic order of the source or target language
(3) paging through successive multiterms beginning with or containing the query term.

The above search operations can be made more sophisticated using 'intelligent' search techniques eg if a Uniterm exists as part of a multiterm only, then the computer should be able to find it. If no match is found, interaction with the user may take place eg the computer may prompt the user to supply a synonym, and then carry on the search with this new information. Manual or automatic morphological truncation

of terms will also prove useful, in case, for example, a term is input in the plural form, when the stored term exists in the singular. The major aim of introducing 'intelligent' searching is to ensure that the computer carries out an exhaustive search, and that even when this fails, it is able to be as helpful as possible, by offering related and relevant information.

Output Formats

Given that output formats are dependent on the needs of individual users, it is proposed that fully parameterized output options be offered. That is, users should be able to choose not only which information they want, and in which form, but also such details as eg page coverage, line spacing, number of columns, character set, type of 'package', and so on. As operations on an open set of options are involved, only a few possibilities are mentioned:

Two basic types of information are usually required by users:

(1) a complete term record, or selected fields thereof, perhaps 'packages'; (this is typically for online use)

(2) selected fields of more than one term record, output in the form of, for example:

monolingual alphabetic indices	egs	term + generic term
bilingual indices		term + synonym + translation equivalent
text-oriented glossaries		term + synonym + source + translation equivalent
alphabetic/systematic glossaries		(many other combinations possible)
by subject area(s)		
by language(s)		
by project(s)		
by source(s)		
etc.		
phraseological glossaries		
concordances		
keyword indices		
full-scale dictionaries		
etc.		

Output Media

In order to provide a wide range of services to a wide range of users, the following output media are necessary:

—VDU or visual display unit. This has the advantage of offering great versatility.

For example, one may receive anything from screenfuls of information to single lines. Screen 'windows' may be employed. A translator could have one section of his VDU screen reserved as a working space, another for calling up information from the LDB. Various types of terminal exist, such as 'slave' terminals, which are directly connected, and have no processing capability of their own, or 'intelligent' terminals, which, as the name suggests, have a certain amount of individual processing power. One may of course wish to connect one's own office or home computer to the LDB, via a

telephone link. It would also be possible to work totally independent of the LDB (see below), using one's own computer.

—Hard copy. A variety of printers should be available, to provide various degrees of quality output, which could be supplied on various types of paper. A user should also be able to receive information on his own printer. Updated printouts of eg glossaries could be sent on a regular basis.

—Microfiche. Advantages here are low cost and regular updates. Such media are very useful for infrequent but detailed searches. Also, it may be the case that some users may prefer to work with hard copy or microfiches to begin with, especially in the early development years of the LDB, and only acquire conversational capability at a time when the LDB can supply a useful number of responses in relevant subject fields.

—Magnetic tape. Such a medium may be useful for eg publishers, who wish to submit many thousand terms for processing. Tapes will also be used for exchange purposes with other LDBs and terminology centres.

—Floppy discs. Subsets of the LDB could be recorded on floppy discs, for use in the user's own computer system. This allows the user to work independently of any link to the LDB. The advantages here are again low cost and the possibility of regular updates.

The accompanying figure shows a possible configuration of a British LDB. (Notes: DB = database; OCR = optical character recognition device; COM = computer output microfilm).

Advantages of using a LDB

We have already mentioned several advantages of a LDB in passing. Here I would just like to draw your attention to the advantage of using a LDB over looking in a dictionary, or consulting a subject specialist, two of the most widely used methods of solving a terminological problem.

With a dictionary, you may find that even a recently published edition may be out-of-date. Consulting a subject specialist may be fruitless, as he himself may not be aware of the term.

A dictionary is time-consuming to use (especially if you share one, and someone else is using it) and consultation may likewise be time-consuming (especially if your specialist has gone for coffee, or you enter into a conversation).

With a dictionary, you must know how it is organized, in order to be able to use it efficiently. Consultation may involve lengthy explanation of the context, or description of the conceptual environment of the term.

When searching in a dictionary, one is usually confined to a 'main entry' type of search. Also, the dictionary, being printed on paper, is of fixed format, and so may not be suited to your especial needs. Other disadvantages of a dictionary are that it is bulky, prone to wear and tear, and not particularly cheap. Failure to find a translation equivalent for a term, for example, whether as a result of dictionary look-up, or of consultation, may encourage creation of a neologism, which in many cases may hamper communication, rather than aid it.

When one looks at a LDB, however, the following advantages are immediately apparent to the professional linguist. The LDB's terminological data are up-dated constantly, with new terms being inserted, obsolete terms excised, and new information being inserted on existing terminology. Access to the LDB can be very rapid, if working on-line, or with one's own subset of the LDB on floppy disc, or with a microfiche or printout tailored to one's needs. The LDB can be considered as a 'black box': the user does not need to know how the data are organized inside the machine. He is helped in his search by powerful search routines, whose existence he is again unaware of, as his contact with the LDB is through a query language which is constructed so as to be as 'user friendly' as possible, and may in fact represent a restricted subset of his own language. These powerful search routines mentioned ensure that a search is as exhaustive as possible. The computer can carry out parallel searches in a number of different data bases, in a number of different term records, and so on, comparing, correlating, combining information in order to produce not just a correct response, but, in the case where a primary search proves negative, a response that may go some way to providing the user with at least some information regarding his query. Exhaustivity and reliability, combined with the authority of a well-run and widely respected institution, will ensure that the LDB offers practical and useful services to all its users.

Costs to user

At this stage, no decisions can be taken regarding costs to the user. However, analysis of practices in other LDBs suggests several methods of payment for services. It is to be hoped that a British LDB will offer a combination of these, suited to individual users' needs. Methods employed in other LDBs are:

Sponsorship system. This would involve an annual grant in return for free use of the LDB, a system practised by NORMATERM.

Subscriber system. This would involve for example a monthly sum giving credit at special rates, again a system used by NORMATERM.

Ad hoc system. This involves payment on a time or unit basis, and is a method practised by all LDBs.

Contributor system. Supplying data free of charge against payment or in return for services is a system offered by eg TEAM.

Partnership system. This would involve supplying data in return for credit to use the LDB, and is a system practised by TEAM and TERMDOK, with a great deal of success.

Conclusion

Most of the services of the LDB would be non-competitive, as they would not be available in any other way. On the other hand, users will consider paying for these services only if they lead to a reduction in their own costs, if they represent a necessary improvement in the quality of their work, ultimately reflected in greater income to justify this expenditure, or if they contribute in some other way to increased productivity, new products or services. If for example the job-satisfaction and productivity of translators can be increased by, say, 10%, the translator, or his employer,

may feel justified in spending an equivalent amount on LDB services. However one looks at it, it would appear to be the case that the demand for translation, and thus for the services of the LDB, will be increased, due to the existence of the LDB.

I stress that what has been presented here is one of several possible models for a British LDB. However, we believe that this particular model is best suited for a British market. This conclusion has been reached due to wide consultation of potential users and contributors. Whether a British LDB is eventually set up or not, is not a decision that we who carried out this feasibility study are empowered to take. What may definitely be said at this stage, is that a large body of interest exists, that this interest is manifested in many different sectors, and it is this interest which will proclaim not only that a British LDB could exist, but also that it should exist; not only that it could be used and supported, but also that it will be used and supported, for the design that has been presented to you this afternoon has been based first and foremost on the needs and expectations of the person who will benefit most from it: the user.

REFERENCES

1 ARTHERN, P. J. (1978), 'Machine translation and computerized terminology systems: a translator's viewpoint' in SNELL, B. M. *ed. Translating and the computer.* Amsterdam: North-Holland, 1979.

2 SAGER, J. C. *and* MCNAUGHT, J. *Selective survey of five Linguistic Data Banks in Europe.* Report prepared for the British Library. Contains a comprehensive bibliography (165 items). Manchester: CCL/UMIST, 1980.

3 SAGER, J. C. *and* MCNAUGHT, J. *Specifications of a Linguistic Data Bank for the UK.* Report prepared for the British Library. Manchester: CCL/UMIST, 1980.

4 SNELL, B. M. *ed.* (1979), *Translating and the Computer.* Amsterdam: North-Holland, 1979.