Krista Varantola
University of Turku

Term banks, text banks and bank users

I am going to take a very concrete, user-oriented point of view in this article. I shall restrict the user-orientation to refer to special field translators and their term bank or rather electronic data bank needs within special fields.

Background.

The Centre for Technical Terminology in Finland, TSK (Tekniikan Sanastokeskus) made its term bank, Tepa, available to the general public in the summer of 1987. The bank is directly accessible. Users, whether public institutions, private companies or private persons, sign a user contract which is free of charge. The only charge is for the use of the computer time. The rate is at present 240 FMK/h.

The bank is situated in a Helsinki University of Technology computer and is accessible 24h/day, also at weekends. The normal user prerequisites are thus a PC and a modem.

At the moment, the bank is Helsinki-centered in the sense that its use is much cheaper if you happen to live in the Helsinki area. For users outside Helsinki the telephone costs can easily become a factor prohibiting the use of the bank. In addition to normal telephone lines, it is, however, also possible to use various data transfer networks that make the telephone bill considerably smaller. I shall return to ideas of decentralisation later.

The bank is now going through its introductory phases and the number of users is still relatively low but increasing fast.

The bank consists of three data bases, Tepa1, Tepa2 and Vniiki which is a 1600 record sample of a Soviet data base consisting of GOST standards. The cyrillic alphabet has been transcribed in all data bases. The bank has now (December 1987) around 12 000 records but the numbers will double in the near future.

Tepa1 consists of a glossary on fire prevention and a glossary on labour protection. It has the highest quality rating of the

Term banks, text banks and bank users
Krista Varantola
Proceedings of NODALIDA 1987, pages 301-311

three data bases. The records come from glossaries that are the result of the co-operation between subject specialists and terminologists. In other words, the glossaries adhere to terminological principles and contain definitions and explanations that give the user an idea of the conceptual systems behind the terms.

Tepa2 is based on subject specialist work but the methodology varies and the terms often lack definitions.

The glossaries included cover such areas as welding, foundries, process automation, clothing, road and transport and forest economics.

Of the future planned data bases, TEPA3 will consist of brand new term records not yet available in printed glossaries and TEPA4 of various types of information that has been handed over to TSK.

In other words, due to pressure from the user end, various types of information are being made available but it is ultimately left to the users to assess the reliability of the information they obtain. The users' task is, however, made easier by having a well-organized bank, by using different data bases with varying reliability estimates. The bank is also interactive and users can send their comments to TSK, through the mail box of the bank.

To summarize, it could be said that the bank is not ideal but it is good enough to be useful. If we think of the slowness and costs of high quality terminology work and of user needs on the other hand, some kind of compromise must be found. It seems very sensible to make public a well-sorted small bank that still has the flexibility of reacting to user comments and changing the approach if necessary.

The project.

TSK is research-oriented and keen on co-operation with university institutions. It was thus agreed that we at the School of Translation Studies, University of Turku, should make an analysis of the term bank as seminar work and from the special field translator's angle. The preliminary work, getting acquainted with the technical aspects and defining the aims of the project has begun.

We have at our disposal a PC with a hard disk (IBM compatible), a 1200 baud modem and main frame capacity at the university computing centre. To access the term bank we do not use the normal telephone lines but the universities' data transfer network FUNET which we access through the university's mainframe DEC20. In the analysis it may also become necessary to use the university's IBM mainframe to run certain textual analysis programs. These technical aspects mean that the time taken to learn the various tricks of the computer trade is not inconsiderable. Enough time must be allocated for learning to use the equipment and utilize its potential. This is a relatively new phenomenon within the humanities but very much a fact of life that should be seen as necessary methodological training also in language and communication studies.

The aim of the project is to study the usefulness and future potential of the bank from a special field translator's point of view.
We intend to study the bank from a theoretical structural angle - analyse the instructions given to users, assess the concrete value of the various record types and use case-studies, i.e. think in terms of hypothetical translations from Finnish into a foreign language within a field present in the bank.
The software we intend to use consists, naturally, of the search routine of the term bank, the TRIP program, but also of text analysis and concordance programs available for PCs and the mainframes. With the content-analysis programs we hope to find out about the lexical and textual structure of the records in the bank and of other special field texts that we use as back-up material.

Problem-setting.
Accessing. What are the optimal ways of accessing a term bank? It is certainly a very good thing that TSK has decided to open the term bank to the public with direct access and without intermediaries. Its location in Helsinki, however, poses a problem for outside users. Telephone costs may become high in the long run if you want to do thorough searches or become bank-addicted. This is obviously not a problem for big users but

certainly for small-scale users. The computer time, on the other hand, is surprisingly cheap. (A 40 minute search which gave over 2 000 chargeable segments cost about 13 FMK).

Alternative approaches might be worth thinking of and we are therefore going to ask what kind of benefits could be derived from a multi-channel access system.

Even a human answering service might prove useful. An experienced intermediary could deal with questions coming in either electronically, by telephone or by post depending on customer needs, equipment and attitudes.

Other questions that will be asked are:

Would it make the bank more flexible if it were accessible in a number of computers in Finland ,e.g., at universities other than the university of technology in Helsinki?

Should the customers be able to buy parts that they need frequently on diskettes? Could an updating service be combined with the diskette service? Should parts of the information also be available in portable form or in a form that could be transferred into a portable electronic dictionary or handbook?

Many advantages are obvious in a multi-faceted service distributing terminology and special field text data. Future translators will rely more and more on electronic data that is directly accessible during the translation process by means of refined search routines.

A translator's work station will have (at least) desktop publishing facilities, consist of a powerful PC and a CD ROM drive, with word processing, electronic glossaries, dictionaries, enclyclopaedic information etc.

Context-oriented approaches such as REFTEX, based on syntactic and textual information available in bilingual parallel texts will certainly prove very helpful tools (cf. Kjærsgaard, forthcoming).

The worst bottleneck for individual, tailor-made data collections will be overcome when a reasonably priced scanner becomes good enough for reading in textual data from sources typeset in various ways.

Various structured special field text corpora will also be available in PC format for a number of referential purposes.(cf. e.g. Brekke, forthcoming). Commercial data banks and free

general services also provide a wealth of special language information useful for translators, on an international or national basis (e.g. the data banks accessible through Easynet, the Videotex services in Finland, France's Teletel etc.)
So, what could a central term bank's answer be to the changing concept of special field translator or communication work?
Before speculating on answers we can look at the question from the customer's point of view.

The LSP translator.

The main problems for a non-specialist LSP translator can be summarized as problem finding and problem solving. In the training we can do our best to increase the future translator's awareness of the hidden problems of LSP texts. For example, we can emphasize the rules of expert-to-expert communication, the demands of certain strictly controlled text types such as contracts and patents, the demands of manuals, instructions, text books, popularizations etc., in other words, of various types of special language communication situations (Cf. Varantola 1986 and Varantola, forthcoming).
We can also make the students aware of their lack of extralinguistic, encyclopaedic knowledge in a special field by drawing parallels from general knowledge inference patterns and associations. We can, for example, ask what kind of associations, public and private, we automatically obtain from pieces of information.
If we start a Finnish short story by saying

- It was January, smoke from the chimneys rose straight up towards the clear sky, the children's hour was about to begin. I decided to have a swim ...-,

our public associations tell us that it was cold; also, at least if you are a parent of relatively small children, that it was about six o'clock p.m. and therefore dark; the television channel in question was number two; the swim would have to take place indoors.
The private associations might include wondering whether the person involved has a private swimming pool and if so

assumptions of affluence or, on the other hand, thinking that it is the hour when the children terrorize the whole house.
If, however, the context were different, a particular technical field and the text run

- The headbox has been greatly simplified, with no internal showers, fewer rotating parts (i.e. removal of rectifier rolls and their drive)....
The sheet is essentially formed and drained at the suction breast roll, and the deflectors under the wire serve simply as support. -

our associations would be very meagre and mainly based on whatever textual cues we would obtain, unless we were experts on paper machines.
What I am driving at is that the problem-solving stage within a special field is where a central term bank can play its part and help a translator to build up the conceptual framework of the field, to cope with the associations, the information hidden in the covert inferences. (Cf. ARK 36, 1987 where the structure and contents of a multi-purpose and multi-language legal database is discussed in detail.)
We must also keep in mind that we are mainly thinking of translators who most of the time translate into a foreign language and not into their native language. It is a fact of life that can hardly be changed in the case of small languages and which adds to the types of textual information that would be helpful in special language translation.

A multi-channel, multi-base data bank

What kind of term bank, or should we say LSP data bank information, would be most useful? What is not there that should be there? How important is the quality vs. quantity distinction?

The central term bank could have, within its field of operation, the overall function of a clearing house. In this capacity it would control the information flow, the information it collects, receives or knows about, and distribute it in various electronic ways suggested above. (Copyright and resource issues are simply

disregarded in this context).
The term bank function is naturally a basic one. A bank, regularly revised and updated, is an indispensable tool in all special field communication. With this basic function under control, it is, however, possible to build some variety into it.

It has already been pointed out that multiple formats of the information would make the utilization of the data very flexible.
From the point of view of comprehension, it is the definition and explanation parts in term records that are particularly useful for translators. The definitions delimit the range of the concepts,i.e. exercise the normative power of terminologies by trying to control the indeterminacy of meaning which is such a basic quality of general language (cf. de Beaugrande, forthcoming). The explanations, commentary parts (which could be far more elaborate) also give very valuable information about the conceptual framework for the uninitiated. They relate the concept to other closely related concepts, warn of misconceptions or misleading usage, or discrepancies between the conceptual systems in different languages. This type of information builds up the insufficient extralinguistic knowledge which is most frustrating for the translator and a major source of mistranslations.
Another way of building up extralinguistic knowledge in a dictionary context is to give special field background information in the same fashion as is done in such general dictionaries as **Collins English Dictionary**:

Planck constant or **Planck's constant** n. a fundamental constant equal to the energy of any quantum of radiation divided by its frequency. It has a value of 6.6262 x 10 joule seconds. Symbol :h. See also **Dirac constant**.
There is also an entry for Max Planck:
Planck Max (Karl Ernst Ludwig). 1858-1947, German physicist who first formulated the quantum theory (1900): Nobel price for physics 1918.
This particular example was chosen because it was among the items in a quiz that was published as part of a **Newsweek**

article (April 20, 1987). The topic of the article was the type of associative knowledge literate Americans should have at high school level. What they were supposed to know about Planck's constant, according to the quiz, was that it is a quantum theory cornerstone. In other words, the associative knowledge required is not very deep, on the contrary, rather superficial but enough to keep the reader on the right track. And this is the type of knowledge a translator could also benefit from in a special field context. The background information need not always be thorough and detailed. A few helpful hints will do the job in many cases when in doubt about the correct alternative or interpretation.

Another question is, how the hints should be offered; as small-scale "cue" entries in a term bank or large scale encyclopaedic entries on a compact disk or a text bank? (See ,however, Svensén 1987:157 ff. who does not think that encyclopaedic dictionaries are very useful.)

Yet, the advent of CD ROM encyclopaedias and dictionaries, etc. does not do away with the need of more "controlled" text bases in special fields.

We can imagine that the need for terminological, i.e., lexical information and certain types of extralinguistic knowledge could be catered for by the expanding term bank, a bank that is in a constant process of updating and revision - but what would take care of the phraseological, syntactic and textual needs of the translator?

It is obvious that filling in all kinds of standard text formulas, e.g., in contracts, patents, specifications, instructions, etc. will be fast and routine if they can be found in a database.In addition, translators would also benefit from a wide selection of controlled text types and sample texts, that could be searched with sophisticated programs. In this way, translators would have at their disposal, in electronic form, parallel texts that they now mostly have to find manually (cf. however Kjærsgaard, forthcoming and the REFTEX system). Finding suitable parallel texts tends to take a great deal of time and when they are found the search procedures are not very sophisticated. With automatic search it would be possible to find differnt types of information, verb+noun, adjective+noun

collocations, use of adverbs and of prepositions, often a tricky business at least for Finns, ideas of word order, sentence structure and textual patterns. (See e.g. Picht, forthcoming). If the texts included in the base represented a wide range of text types with varying degrees of LSP-ness it would be possible to make sophisticated comparisons of term depth, strictness of term use, term density (use of fillers to control the rate of information flow), proportions of covert and overt information, grammatical features, etc. (cf. Varantola, forthcoming).

Information of this type would add to the translator's intuition of style and successful LSP communication. And, this kind of "acquired" intuition is essential in any type of textual modification task.

To speculate on the nature of the texts, it would seem sensible to include the texts from which the terms have been excerpted, but other, up-to-date texts would be welcome, too. A central LSP databank could certainly benefit from the knowledge of individual translators who, over the years, have collected valuable private term and text files that could be made more generally available through the bank. If there is no way of checking and guaranteeing the reliability of a private or a company collection, it should still be made available with the necessary reservations, because top quality terminology work is extremely time-consuming. I do not think there is any need to hold back information if its expected status is clearly indicated.

In other words, quality is the key concept but the need for quantity and versatility is the pragmatic demand that has to be satisfied. And, there seems to be little reason why the two seemingly opposite demands could not be combined. Basic research into text selection and ideal term record structure is essential. New findings will in due course be applied in practice but, at the same time, an existing bank has to react to customer demands. If the bank is well-structured with clearly differentiated data banks and not a Christmas pudding type mixture, it should be possible to maintain it in a constant stage of revision and development. Up-to-dateness and flexibility are after all the basic advantages of an electronic data bank.

It seems, however, that the term and text bases should remain separate but mutually compatible, e.g., in the sense that you would know that each term present in the term bank would also be available in a text context.

A clearing house of the type sketched above would naturally involve an immense amount of work. On the other hand, there is no reason why it should be done on a small scale. Co-operation between LSP data bases is necessary,as is exchange of material and the use of other data bases compiled for other than LSP term- or textbank purposes. General services of the French leletel type and others mentioned above, could prove selectively valuable.

It is also clear that although I have taken a narrow LSP translator's view, other groups of users, subject specialists naturally, but also industry in general, professional training, LSP research, etc. would benefit from a multifaceted system and could also contribute to it.

In this article I have tried to bring up ideas that we hope to study empirically during our small-scale project. It is an advantage that we have a chance of this hands-on approach at this stage when the bank is still small and easier to conceptualize than a large bank.

REFERENCES

ARK 36. 1987. Sproginstitutternes Arbejdspapir. Handelshøjskolen i København. Copenhagen.

de Beaugrande, R. Forthcoming. "Text as the new foundation for linguistics". Paper presented at the VIth European Symposium in LSP, Vaasa 1987.

Brekke M. Forthcoming. "The Bergen English for science and technology (Best) corpus: A pilot study". Paper presented at the VIth European Symposium in LSP, Vaasa 1987.

Collins Dictionary of English. 1979. Glasgow.

Kjærsgaard, P.S. Forthcoming. "REFTEX - A context-based translation aid". Paper presented at the 3rd Conference of ACL European Chapter, Copenhagen 1987.

Newsweek. 1987.(April 20). "A Dunce Cap for America. What we don't know, why we don't know it".44-46.

Picht, H. Forthcoming. "Fachsprachliche Phraseologie". Paper presented at the VIth European Symposium in LSP, Vaasa 1987.

Svensén, B. 1987. Handbok i lexikografi. Stockholm

Varantola, K. 1986. "Special language and general language". Unesco Alsed-LSP Newsletter, 9:2. 233-244.

Varantola, K. Forthcoming. "LSP translation from a linguistic angle with special reference to engineering texts". Paper presented at a symposium on translation and LSP , University of Surrey 1986.

Varantola K. 1987. "Popularization strategies and text functional shifts in scientific/technical writing". To appear in UNESCO-Alsed LSP Newsletter.