[*Translating and the Computer 9*. Proceedings of a conference… 12-13 November 1987, ed. Catriona Picken (London: Aslib, 1988)]

The LEXIS termbank

Erika Hoffmann

Bundessprachenamt, Hürth, West Germany

THE NATURE OF THE MATERIAL INCLUDED

LEXIS, now in its twenty-first year of existence, was designed right from the outset as a database to be queried for or by translators, i.e. technical translators who specialise in one or more fields. Therefore, the records stored in LEXIS are predominantly LSP terms (language for special purposes), mostly compounds, while LGP terms (language for general purposes) are represented to a lesser degree.

The data record is a language pair with a foreign language (of which we have eight: English, French, Russian, Portuguese, Dutch, Italian, Spanish, Polish) and German being either the source or the target language. This reflects the translation work being done at the BSprA: from a foreign language into German or vice versa.

Apart from the language pair, the data record contains a code group consisting of the language symbol, subject field code, source code and quality symbols, as well as optional short comments in brackets.

The vocabulary of LEXIS covers the fields of technology, natural science, medicine, military applications – to give just a few examples – and is partly standardised by our terminological committees or by national or international standards institutions. We have approximately 220 subject fields (developed from the broad fields of UDC) of which electronics, electrical engineering and data processing are well covered in the main languages, i.e. English, French and Russian. In contrast to termbanks with multilingual (meaning more than two languages) data records, it is not necessary for us to increase the vocabulary of the 'lesser' languages to the same volume as the main languages, one reason being

the heterogeneity of subject fields in the texts which vary widely depending on the language from which or into which translations are done. The most striking example is the vocabulary in Polish which reflects unusual subjects ranging from Silesian refugee meetings, family reunions, German and Polish pension laws, the Nazi past, etc., to the peculiarities of Eastern European economic systems. For example: eksport wewnetrzny — Binnenexport or 'internal export', meaning the sale of domestic goods on the domestic market for hard (Western) currencies, a phenomenon not normally to be found in the Western world.

In this connection, I should also like to mention the situation in which we find ourselves, namely of having two German languages (GDR-German and FRG-German). When translating texts concerning Eastern Europe from Russian or other Slavic languages, but also from Western languages, into German we are supposed to use 'Federal German' terminology. However, where there are no equivalent concepts, we use the terminology of the GDR, which we mark as such.

Due to the fact that — on historical grounds — our data record has a fixed length, it became necessary to accommodate text information such as definitions, context, etc., in a second storage, called background storage, which became operational in 1983. This storage is monolingual and comprises the eight languages mentioned above plus German as a separate language, which corresponds to practical usage: one rarely finds congruous definitions in two or more languages (consequently, you will mostly find *translations* of definitions, e.g. in ISO). We found that only one-third of the entries require definition as most compounds are self-explanatory.

THE WAY IN WHICH TERMINOLOGY IS COLLECTED

A considerable part of our terminology is the result of feedback from translations in the form of word lists containing queries. These lists are completed by the translators (as far as they are in a position to do so) and checked by the editor when verifying the translation.

The terminologists process these lists by delimiting the terms (they are mostly compounds, as mentioned, and must be stripped of any accidental additions, such as adjectives, etc., that do not belong to the term proper), checking them against the database, harmonising them with existing entries or parts thereof; they try to find more information about the term in reference works or literature, and contact the translator and editor if further clarification is required. After the code group and comments have been added, the terms are checked once more by the section where they originated, and are input. The database is updated at regular fortnightly intervals. The data for the update is entered continuously, placed in intermediate store, verified, released and included in the database after several checks have been carried out by the system to avoid formal errors and duplicates.

Technical publications are the second source of terminology. The terms are processed in the same way as described above and approved by a section that is competent in the particular field. This type of terminology collection is rather time-consuming as one rarely finds parallel articles in two languages, i.e. articles which are originals and not translations.

HOW IT IS USED IN PRACTICE WITHIN A WORKING ENVIRONMENT

The translators type the queries in the form of a list and receive the printout with the answers in the afternoon of the same day if the queries are input by 1.30 p.m. The background storage is queried automatically for definitions which are output with the other queries. The advantage of a printout (as opposed to a screen display) is obvious: the translators have the answers in black and white, they can complete the list and suggest corrections; when adding new information they can indicate the source (an expert, a reference book, etc.), thus making it easier for the editor to verify the terminology used in the translation. In urgent cases they can call a terminal or go to it: the queries are input in the interactive mode and they are given a printout of the screen.

All the translators have a set of microfiche and a reader. They use them frequently, especially if they want to scan an entire field of terms starting with the same elements. In this way they can often construct the term they are looking for by analogy. This method is convenient for English and German compounds as they are structured in a similar way.

Within LEXIS, a subbank, which we have named ‘Register’, was established to accommodate ad hoc glossaries for special purposes. Here, a record may have from one to three columns of 120 characters each for linguistic data and a fourth column of twenty characters for additional information. The requirements are not as stringent as they are for LEXIS data records. This subbank can be updated continually, independently of the LEXIS update, and the vocabulary can be re-sorted as often as necessary. It has proved invaluable for trilingual glossaries (for our tripartite projects which are mostly English-French-German), for glossaries prepared during the translation process by the editor or editors when several translators work on the same text, or any kind of word lists that have to be updated frequently and arranged in alphabetical order. The upper limit is 10,000 data records for each glossary. At present these data records do not automatically become part of LEXIS. An automatic conversion is envisaged for the not too distant future.

PROBLEMS IN MAINTAINING THE TERMBANK

Lack of Feedback

There is hardly any feedback from translators who use the microfiche.

If translators do not query the termbank by list, they are unlikely to put down the terms they have found. With a list, there is more incentive to contribute terminology quantitatively and qualitatively, e.g. by making suggestions to add new and to correct existing terms.

Dictionaries are taboo

Dictionaries, standards, etc., are protected by copyright and contain a warning: 'No part of this book may be reproduced in any form without written permission from the publisher'. When we asked for such permission some time ago, we received a bill from the publisher charging an amount which we were unable and unwilling to pay. We have never tried again. We do not copy dictionaries, and in the case of definitions which, for us, are hard to come by, we make it a point to use only official publications which we are permitted to excerpt and for which we do not have to pay a fee. We have stored the definitions from the AGARD (Advisory Group for Aerospace Research and Development) dictionary which had the advantage of being supplied on data medium (magnetic tape) so that we could feed them in directly. Similarly, we input definitions from our manuals and regulations directly from computer tapes that were produced for printing the manuals.

Quantity versus quality

During a recent audit of civil service posts we were required to specify the quantity of terms a terminologist collects and processes. The underlying idea was to find out whether existing posts were sufficient and whether new posts should be provided or existing posts reduced.

We were unable to quantify the performance of each terminologist since we felt that it is quality that matters rather than quantity. As LEXIS has reached a phase where data must be purified to adjust it to current usage, the growth of the termbank no longer reflects the work that goes into maintaining it. Furthermore, our system does not provide for an indication of the originator of the term (the terminologist). As soon as the term is included in the database, there is no way for the system to find out who the terminologist was. Hence no statistics relate to individual terminologists. All that we were able to do was to take the statistics automatically maintained by the system which counts new entries as well as deletions and modifications of existing entries, and divide the total number of pertinent computer operations by the number of terminologists; the result was twenty-five operations (new entries, deletions and modifications) per work day and per terminologist.

The danger inherent in such statistics is that the emphasis is shifted from quality to quantity, which should be avoided at all costs.

Selecting the terms for storage

A very important question for each termbank is to decide whether to collect and store everything that it comes across or to make a choice. Since there is practically no limit to the storage capacity of modern systems, one might be inclined to 'save everything' and then to decide at a later date what to leave in the termbank and what to delete. It has been our experience that terminologists only reluctantly part with a term once it is stored.

Hard as the decision may be, it is wise to make a selection of what to store. It is not advisable to make a decision at a later date, since the origin, author, or context, etc., are no longer available. We invested a great deal of time and effort into finding German equivalents for French neologisms, but then decided not to continue the project. Neologisms come and go and only very few of them enter general usage. This was one of our 'prospective terminology' projects, but it did not justify the effort that went into it. Another example was cuts of meat in several languages. We decided not to include them, for the chance that they may crop up again is very slight. One solution might be to store such glossaries temporarily in the 'Register' subbank and then decide at a later date what to do about them.

Exchanges between termbanks

Exchanges between termbanks as a means to increase the vocabulary are not recommended for existing termbanks that have a sizeable database of their own. We consider the main obstacles to be:

—the structure of the data record;
—the classification system (subject field key); and
—the difference of the 'user profile', i.e. the vocabularies required by the users in question.

The data record structure is not so crucial if the data record of the receiving termbank has fewer elements and the surplus elements can be disregarded. In the opposite case the missing elements have to be added individually.

The problem of classification is more crucial. The specific needs for which termbanks are designed vary greatly, and so does the classification scheme. The classification scheme for a termbank of a medical institution or a chemical firm will differ considerably from a classification scheme of a termbank like ours which has only six subject fields for medicine and six for chemistry and which would not suffice for a firm like Bayer-Leverkusen or BASF. Our translators admit that the subject field code is only of secondary importance to them. When looking for a target language equivalent, they go 'by intuition' and only when this fails do they look at the subject field code to pick the right equivalent.

A subject field key that is too diversified complicates the work of the terminologists who often have to do extensive research in order to be able to classify a term correctly.

Cleaning up the termbank

Above a certain number of entries it is no longer possible to revise a termbank by starting with the letter A and working one's way through the alphabet, something we tried a long time ago. What we do now is to take an isolated problem, e.g. a term, a spelling variant, etc., and harmonise it with all the pertinent entries whether this element appears isolated, or in the front, centre or end position of a compound. The word 'Stop' is spelled in German with double p, but there are terms where stop is now spelled with one p: Start-Stop-Verfahren. Similarly, 'Plot' has been changed from double t to one t, and 'on-line' should now be written as one word according to ISO. The German 'Fuge' is another case in point. We write Schiffbau, but Schiffsführung, Mitgliedsbeitrag, Mitgliedsstaat, but Mitgliedschaft (without s). There are rules, but they do not always apply 100 per cent. Regional variants and usage in certain branches (the Navy insists on Decksoffizier – deck officer – instead of the correct Deckoffizier, or Saugepumpe – suction pump – instead of Saugpumpe) renders the life of terminologists difficult. The means they have on hand is a concordance which is designed to their specifications, and they try to harmonise the variants – where this is possible – usually deciding in favour of the commonly used form. Often we decide on the basis of frequency or the principle 'of least changes'.

One of the major clean-up operations at present under way concerns cases like 'Kathodenstrahlröhre' which now should read 'Elektronenstrahlröhre'; or 'Schaltkreis' and 'Stromkreis' which should be substituted by 'Schaltung' – but not in *all* cases. This will keep our terminologists busy for quite some time.

A problem which we are unable to solve properly is that of compounds in the romance languages of the type noun plus genitive (noun phrase) or noun plus genitive plus prepositional phrase where the genitive may take the definite or indefinite form: 'trattamento dei dati', 'trattamento di dati' (data processing – Datenverarbeitung), 'data dell'entrata in servizio', 'data di entrata in servizio', (in-service date – Einführungsdatum). Or similarly 'traitement du signal', 'traitement de signal', where the genitive object may in addition take the plural thus giving four entries.

For conventional lookup this problem is irrelevant. However, for database queries, either all forms have to be included or an algorithm has to generate all possible forms. In the latter case, all genitives would be subjected to this routine.

Synonyms are allowed in LEXIS, since the goal of having only one term for one concept cannot be achieved. We offer the translators as many synonyms as correspond to a certain usage; they have to make the choice and adhere to the chosen equivalent throughout the translation. However, from a certain number on we sometimes select just a few (e.g. after consulting an expert) and delete the rest. (Too many synonyms are not in keeping with the principle of standardising terms.) The work of cleaning a termbank is never-ending and should be done continuously.

PROSPECTS FOR THE IMMEDIATE FUTURE

Nordic languages (Danish and Swedish) are to be included in LEXIS, although it is envisaged to store them in the subbank 'Register'.

The number of workstations for translators will be increased and one station will be provided with a graphics capability.

We would like to have access to Eurodicautom via DATEX P, but it is not yet certain whether the required funds will be made available.

RECOMMENDATIONS

Based on the experience gained in twenty-one years of using LEXIS daily, we have come to the following conclusions:

—a termbank should be as 'user-friendly' as possible, hence as simple as possible;

—the data record should not be rendered too complicated by envisaging too many data elements which are nice to have, but difficult to maintain (update). The following elements of a data record should be sufficient:

- source language term/target language term with a minimum of grammatical information (if any)
- definition and/or context
- subject field, source, quality
- status or authority (deprecated, obsolete, DDR, UK)
- date of entry, date of last change

—all activities concerning the maintenance of the database should be concentrated in one place;

—users should not be allowed to establish their own individual database because this would lead to a proliferation of synonyms and it would be difficult to harmonise the various subbases.

CONCLUSIONS

Acceptance of LEXIS by its users is good. We have an average of 1,000 queries per day.

Demand for the microfiche has been rising. The number of copies had to be increased from 3,000 to 4,000 this year.