[From: *Term banks for tomorrow's world: Translating and the Computer 4.*
Proceedings of a conference ... 11-12 November 1982, ed. Barbara Snell. (London: Aslib, 1983)]

STAGES IN THE LIFE CYCLE OF LEXIS

Erika Hoffman
Bundessprachenamt, Hurth, Federal Republic of Germany

ABSTRACT: LEXIS (short for Lexical Information System) is the term bank of the Language Service of the Federal Armed Forces (Sprachendienst der Bundeswehr), which was designed as a machine aid for the translator. Although it has undergone several changes over the years, its aim has remained the same: to assist the translator in his work by providing the necessary vocabulary as quickly as possible and updating the data base at very short intervals.

BACKGROUND

The name LEXIS was introduced in 1975, but the system as such has been in operation since 1966 and was known as "textbezogene Fachwortliste" (text related glossary) although this glossary was only one aspect of the system.

It may be worthwhile going back to the very beginning of the system since prospective operators of data banks might derive some benefit from out experience. I shall go over the first stages very briefly and dwell on the present stage in more detail.

The deliberations that led to the development of the system were as follows: A large translation service wanted to make use of the computer, but did not consider machine translation as an option in view of the low quality of MT output on the one hand and the rigorous demands made on the quality of translation (human) on the other, which means that accuracy and style take absolute precedence over speed and even costs. The way out of this dilemma seemed to be the introduction of machine aids to translation which led to a system that relieves the translator of the burden of maintaining a card index of his own as well as consulting conventional dictionaries, technical glossaries, a central card index etc. In other words, a system to help the translator find the 'mot juste" was the solution required.

FIRST STAGE

The fact that we had some 80,000 English-German and about 40,000 French-German terms in machine—readable form (on punched cards) proved to be a great asset for us. When the system became operational in 1966 (the preparatory work including programming, took two years, from 1964-1965) we had a stock of linguistic data large enough to enter the "production phase" immediately, i.e. we could work with realistic data and not just a small-scale dictionary model. The hardware consisted of the Telefunken computer at Trier, 200 km from our location at Mannheim. The queries were punched on tape and sent to Trier by mail; the lists with the answers reached us 3 to 4 days later. The dictionary was updated on the premises in Trier in the presence of one of us, at more or less regular intervals of about 5 months; new entries and deletions were punched on cards.

After about three years of "commuting" between Mannheim and Trier the second stage was started.

SECOND STAGE

This stage saw our conversion to IBM hardware, which involved the changeover from punched tape to punched card for the queries, and our link-up to the computer centre in Bonn. Teleprocessing enabled up to send out queries via data line to Bonn for batch processing and to receive answers on our printer in Mannheim the same day. The data for dictionary update were sent to Bonn in the same manner. The update itself was carried out at shorter intervals (every 3 or 4 months) and no longer required our presence.

During this stage, at the end of 1971 to be precise, the translation service moved from Mannheim to Hürth to form the Bundessprachenamt (BSprA) together with the Federal Armed Forces Language School from Euskirchen). There were no changes as far as the hardware configuration was concerned, but Russian was added to our data base as the third foreign language.

THIRD STAGE (LEXIS I) (1975 to the present day)

This stage brought us the Visual Display Unit and with it far-reaching changes to our system.

Until then our data terminal had consisted of several card punch units, a card reader and a line reader (all working OFF-line), a printer control unit, modem and a data line.

With the advent of VDUs, the number of ON-line items of equipment increased considerably (the BSprA is just one of the many users of the computing center in Bonn which are served on a time sharing basis) with the result that a new operating system, IMS, had to be introduced at the computation center. This, in turn, imposed several restrictions on our system, but also brought us quite a few advantages.

The limitations of our data records to a fixed length had the greatest impact on our work

40 characters for field 1 (I_1: foreign language term)

34 characters for field 2 (I_2: explanatory addition to or rearrangement of I_1)

45 characters for field 4 (I_4: German equivalent)

39 characters for field 5 (I_5: explanatory addition to or rearrangement of I_4)

13 characters for field 3 (I_3: code group (which was increased by 2 characters): language symbol, subject field code, source, quality symbols)

and I_6 - an 8-digit number allocated (automatically) to each data record as the storage address and for parity check.

Whenever a data record has an element exceeding the number of characters given above it is cut off at this position and stored in full length on tape, where the maximum length is 75 characters for all data elements (except IH3Y and I_6, of course, which have a fixed length shorter than the above limit).

However, the positive aspects of the innovations introduced in this stage outweigh any disadvantages:

- input via VDU with all the convenience inherent in this equipment
- intermediate storage of updated data with repeated output of a protocol for proof reading and correction
- ON-line correction of data records (in the dictionary, prior to this development, even a minor error such as misspelling required the deletion of the whole record and its input as a new entry)
- interactive mode, i.e. presentation on screen of the actual state of the term bank at a given moment
- and, above all, update service at regular fortnightly intervals which we consider to be the greatest asset of our system: the term bank is brought up to date regularly every two weeks.

FOURTH STAGE (LEXIS II) (1981)

As there is no space in our data records for lengthy definitions, context, etc., the idea emerged of offsetting this disadvantage by creating a back-up store. It became operational in the second half of 1981 under the name LEXIS II.

While LEXIS I is basically a bilingual dictionary with German always being either the target or the source language, LEXIS II is monolingual. Access to the data records is via the keyword or the storage address (H_6 is the equivalent of I_6 in LEXIS I). There is ample space to accommodate data elements:

1000 characters each for definitions
500 characters each for context
75 characters each for synonyms, antonyms, broader term and narrower term
200 characters each for literature

There is practically no limit to the number of definitions, context, etc. that can be stored for a given term.
All operations are ON-line. Output is on hard-copy or in interactive mode.

What is the present state of our term bank?

There are three files in LEXIS I and one file in LEXIS II.

LEXIS I

File 1 contains those data records in which the foreign-language term is in first position(I_1) and the German equivalent is the I_4.

File 2 is the inversion of file 1. Every new entry is automatically inverted during the update run. This is, of course, not advisable in cases where the target language equivalent is a paraphrase of the source language concept, e.g. 'kreisfreie Stadt' - independent city not affiliated with a country. A symbol prevents the output of such data records - in this instance in the direction English-German.

The term bank contains (as of September 1982):

864,900	entries	English/German
217,900	"	French/German
319,900	"	Russian/German
6,900	"	Portuguese/German
8,900	"	Dutch/German
12,500	"	Italian/German
750	"	Polish/German (introduced in August 1982)
Total: 1,431,750		

Spanish may follow in the very near future.

The same figures apply to file 2.

The third file within LEXIS I is the compressed or condensed file which was designed for the querying operation. For this file blanks and special characters are suppressed in the data record with the result that the dictionary entry consists of alphanumeric characters only. The queries (or search words) are compressed in the same way in order to bring about a march during dictionary lookup irrespective of whether the term (compound) is written solid, separate, with a hyphen or with special character (slant etc.). Spelling variants (British) are not handled by the system because, in view of the data mass, it would be too time- consuming for all queries to be first checked for possible spelling variants. Since US texts prevail in our translations, American spelling takes precedence over British spelling when an entry is included in our term banks.

The vocabulary of our term bank is alphabetically ordered, it is not hierarchically structured and is in most cases the result of feedback from our translations. The translator queries the term bank. In case of a "no match" the system prints the message "not in the dictionary" (FEHLT). The translator adds the equivalent whenever he is in a position to do so. The list is checked by the reviser as he goes over the translation: he changes, deletes the term or adds some information to is, as he sees fit. The list with the proposed terms is passed to the terminologist who ensures that the terms meet the standards set for our terminology.

He checks the compounds and parts thereof for semantic consistency, grammatical correctness, usage, usage in other languages (which can be accessed from the German equivalent or subject field), adds other information such as subject field code etc.

In case of doubt he contacts the reviser and/or other experts in the field in question. The terms are submitted to the section head (Referatsleiter) for final confirmation.

Once a term has been inputted, verified, corrected and released for update, it is subjected to a doublet control to ensure that identical character strings are not included in the term bank (unless they differ in the subject field). These terms are rejected by the system and printed in a special list. A further list, the "List of Supplements", shows the terminologist which changes were made to the data base on which day. This list is maintained over about six months and then started anew.

The vocabulary is not compiled systematically. We do not copy existing dictionaries. Firstly, dictionaries are protected by Copyright (which we do not like to infringe upon).

Secondly, dictionaries contain terms which are unlikely to come up in our texts. We would consider them to be nothing more than ballast. Our vocabulary is predominantly of a technical/scientific nature and is partly standardised on a national and international basis. For our sector, terminology committees are convened once or twice a year. These committees (12 in all) are specialised in fields such as electronics and communications, aviation and space, optics, automotive vehicles etc. with experts sitting in from industry, Government, institutes etc. During these terminological sessions, which last about two days, approximately 100 terms are approved, usually in the form of a language pair English/German with definition. The terms are prepared by the Secretary and submitted to the members in the form of draft lists well in advance so that they can form an opinion and give their comments during the session. Once a term has been accepted by a committee it is marked in the term bank as such and its use becomes mandatory from that point.

Our entries are compounds in natural word order. We do not group terms around a key word. Inversions and phrases are rare. Apart from marking the infinitive in the English part of a term (to distinguish verbs from noun homographs) and apart from marking the German verbal noun, we do not provide for any other grammatical information as we feel that professional translators should not need this kind of information. At present we have only capital letters and no diacritics (as was the case in the early days of data processing). There are two exceptions: In French we use where necessary the apostrophe as a code to indicate the "accent aigu" in order to differentiate between, for example, "limite" and limité". In Polish, a language in which it is possible to do without diacritics, we use the apostrophe, period and comma.

The correspondence is one-to-one (as is the case in our transliteration system for Russian) so that the vocabulary could be converted automatically to a different representation.

Hardware	IBM 3081 for TP IBM 4341 for batch processing disc storage 2 data lines to Bonn 7 VDU and 3 line printers at the BSprA
Software:	PL/1 and Assembler
Output:	Hardcopy and interactive mode for LEXIS I and II. COM for LEXIS I (i.e. for our branch offices which do not have terminals) Photosetting for dictionaries (LEXIS I)

The next stage is imminent. We are about to install an IBM 8100, which we will use for word processing and as a control unit for out system. Should the interface between word processing and data processing prove to be reliable we might consider it for the output of smaller glossaries.

As a next step we envisage for our term bank:

- the use of upper and lower case
- the use of diacritics

} conversion of the data base to this form of representation will involve a great deal of tedious work

The introduction of a variable length of data elements in place of the present fixed word length.

We do not consider our System to be as sophisticated as it could be, but

- it is easy to operate
- it is very fast
- and it offers the translator only as much information as he can handle and requires for his work, i.e. the foreign-language term and its German equivalent(s), subject field code, source, quality symbol and, on demand only, definitions and context.