Seija Suonuuti, Pertti Nuutila, Nokia Telecommunications

# Ways and methods of utilizing a termbank

*ABSTRACT: This paper deals with some of the ways in which a termbank can be used. The paper covers two slightly differing subjects. The first part concentrates on the contents and structure of the termbank at Nokia Telecommunications while the second part deals with the application of the termbank to a machine translation system.*

## 1. Termbank TERO - basic structure and traditional use

TERO is the termbank of Nokia Telecommunications. The main emphasis in the construction of the termbank has been on creating a reliable database that covers the terminology of our research & development work in telecommunications. Therefore the majority of the terminology in the termbank is the product of various terminology projects. Additional material has been included as a result of standardization follow-up studies and from the feed-back received from the termbank users.

### 1.1 TERO in a nutshell

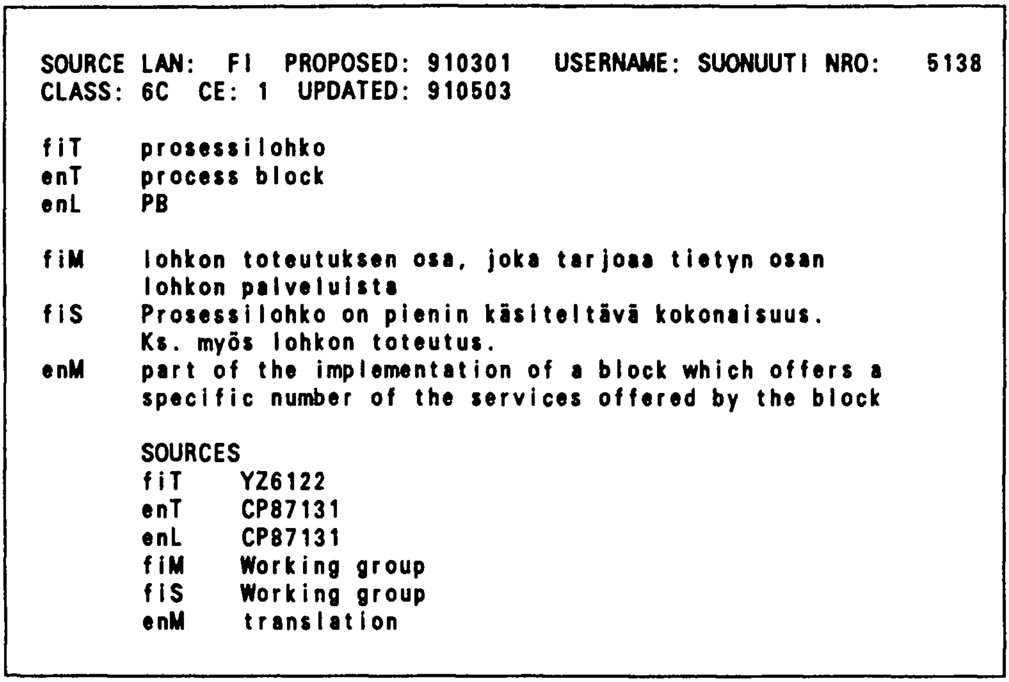
The following figure displays the information stored in the termbank.

| | |
|---|---|
| languages | Finnish, English, Russian, French, German, Swedish, Spanish, Portuguese |
| term information | terms, deprecated terms, abbreviations, mathematical symbols |
| definition information | definions, notes on the concept or term usage, comments for project handling |
| additional information | e.g. classifications, source information, creation date, update date |

Figure 1.1. Information stored in the termbank.

### 1.2 The structure and management of the termbank

Each field contains only one term, abbreviation, symbol or definition for one concept. The following figure is a hypothetical example of one term record.



```
SOURCE LAN:  FI  PROPOSED: 910301   USERNAME: SUONUUTI NRO:    5138
CLASS: 6C  CE: 1  UPDATED: 910503

fiT    prosessilohko
enT    process block
enL    PB

fiM    lohkon toteutuksen osa, joka tarjoaa tietyn osan
       lohkon palveluista
fiS    Prosessilohko on pienin käsiteltävä kokonaisuus.
       Ks. myös lohkon toteutus.
enM    part of the implementation of a block which offers a
       specific number of the services offered by the block

       SOURCES
       fiT    YZ6122
       enT    CP87131
       enL    CP87131
       fiM    Working group
       fiS    Working group
       enM    translation
```

Figure 1.2. Example of a term record.

The termbank consists of two databases: the actual termbank database, which is also used for the MT system, and a data base of proposals. The actual termbank database comprises some 5500 concepts.

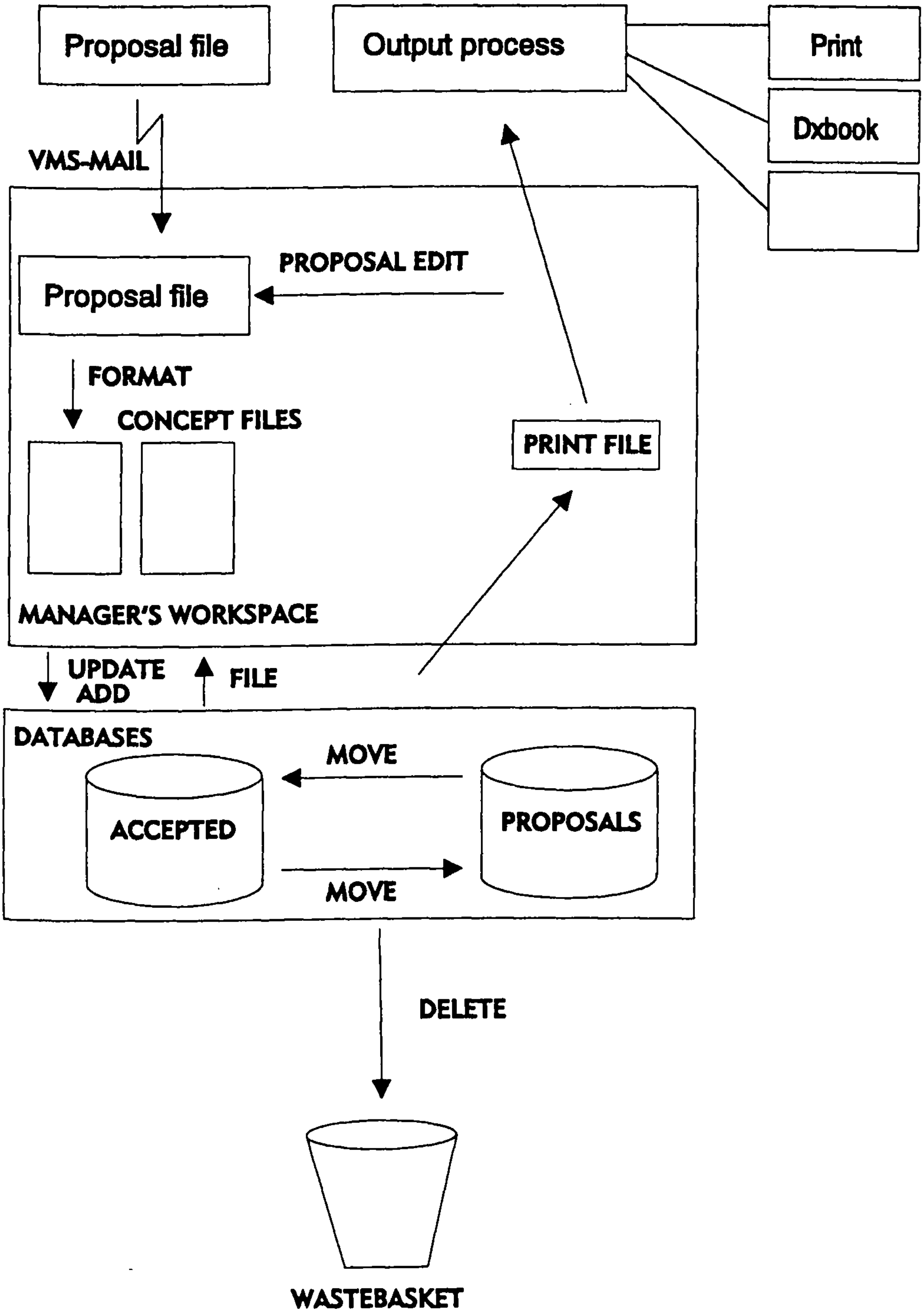


Figure 1.3. The structure of the termbank

The databases are relational and make use of indexed files. The management of the information is carried out in a terminologists' workspace. The smallest unit handled is one concept with a unique concept number.

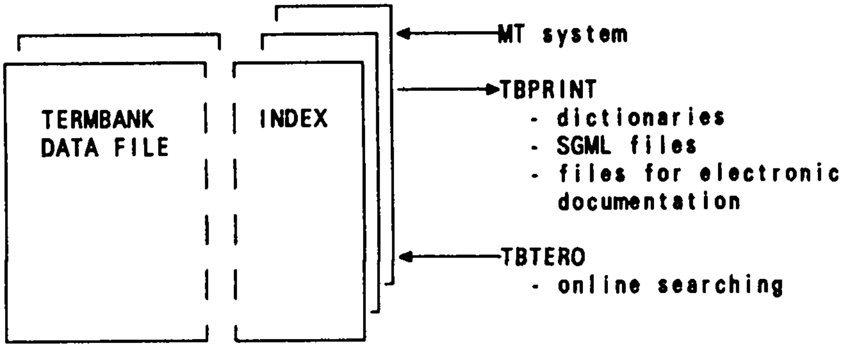


Figure 1.4. Use of the termbank.

The termbank is most commonly used in creating, editing and translating customer documentation. The termbank is also used to some extent for writing system specifications and descriptions. For this purpose a traditional online search program is used.

The material can also be used separately, extracted from the normal termbank system. Two applications have been created for this purpose.

The first application creates files or reports according to a definition file. The output contains terminology for print-outs or some other systems, e.g. dictionary print-outs, SGML formated files which can be included in customer documentation or files for electronic documentation applications.

During the past two years we have also worked on the application of the termbank to the computer aided translation of text. This interface differs slightly from other applications. The main idea being that the termbank is used directly through a specific index instead of a separate report.

## 2. Adapting an existing termbank to a machine translation system

### 2.1 MT system

The MT system is based on the transfer approach and it makes use of dependency grammar. It serves to translate sentences in much the same way as cars are assembled on an assembly line. All processes make use of a number of rules and add information to the words in a sentence. Word selection is just one of the processes on the assembly line and its rules are based on the information accumulated in the preceding processes. Figure 2.1 illustrates how a sentence is processed on the assembly line.

The system is based on a hierarchical lexicon structure. Lexicons are scanned through from the most specific to the most general, and the first occurrence of the word is chosen,

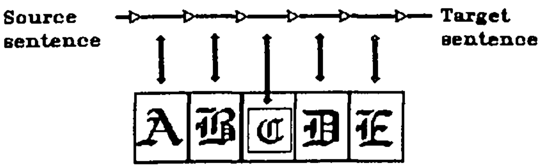


Figure 2.1. Assembly line.

as illustrated in Figure 2.2. In this paper we discuss the most specific of lexicons i.e. the customer-specific lexicon (CS). In our case, the CS lexicon is further divided into three parts according to the sources, which are: our termbank, nomenclature databases, and documents. Here I will concentrate on the termbank.

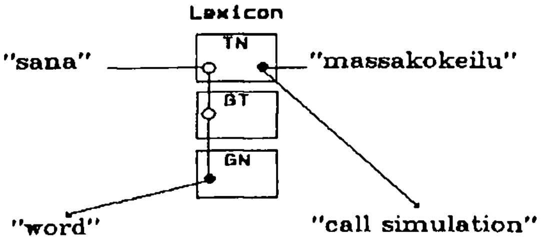


Figure 2.2. Lexicon system.

### 2.2 New structure of the termbank

There are different types of termbanks. Some may contain anything under the sun while others are quite strict in what may be entered. Our termbank is of the latter type. It has been in relatively wide use within our company for a number of years. The knowledge gathered is reliable and is stored in a machine-readable format, and therefore it was only natural to utilize it in the MT system.

The termbank contained the Finnish terms and their equivalents in English, but there was still one drawback – the termbank did not contain any grammatical information. It is obvious that for a lexicon to be used in a natural language system, grammatical information is of vital importance.

It was necessary to make some modifications to the data structure of the termbank to include the necessary data. A special link was made from each Finnish and English term or abbreviation to a grammar table. Similar links may be used if other language pairs are needed.

### 2.3 Handling of grammatical information

Our termbank accommodates words of only four word classes i.e. nouns, verbs, adverbs and adjectives. Nouns are the main class and adverbs and adjectives are quite rare. In addition to these, the termbank contains a number of abbreviations.

Methods and processes for handling the grammatical information were introduced. The actual terminological work is carried out by terminologists as before while translators are responsibile for linguistic work on the grammatical information.

By means of a suitable computer program grammatical information for the terms was added on a default basis after structural modifications. The program assumed that every word encountered was a noun and generated corresponding attributes, e.g. plural form -s or -es. This was first done in batch mode, but in normal operation the program is used to generate default values for all new terms as they are entered in the termbank.

The generated default values must be checked and to serve this purpose a special editor was built. By means of this editor irregular plural forms and other syntactic information is corrected, and semantic information for nouns is entered. In some cases, e.g. when the word class is changed, a new set of attributes must be entered.

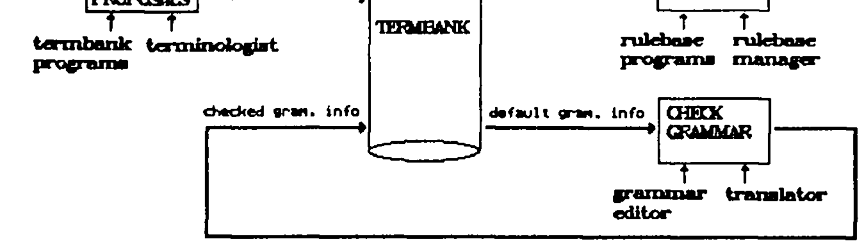
### 2.4 From a proposal to a lexical rule

At present, the process flows as described in the following. After having handled a term proposal, terminologists append the term into the termbank. At the same time a special default value generator automatically appends values in the grammar table, which is an integral part of the termbank.

Translators use the grammar editor to check and correct the generated default grammatical values, and the editor can be used simultaneously by a number of users. This has been accomplished by a locking mechanism. As the user enters the first term to be checked, the program locks the whole concept. The user may then either accept the term record as such or make required modifications. The editor program scans through the termbank and displays all term records whose grammatical information has been labelled with the status "generated".

Finally, the rulebase manager generates the lexical rulebase to be included in the machine translation system. The terms can be selected according to the date of entry or status, and thus it is, for example, not necessary to generate the whole rulebase each time a new term is added. Figure 2.3 illustrates the process from a term proposal to a lexical rule.

Figure 2.3. From a proposal to a lexical rule.



KEYWORDS: terminology, termbank, terminological database, machine translation