[From: *Overcoming the language barrier*, 3-6 May 1977, vol.1 (München: Verlag Dokumentation, 1977)]

AUTOMATIC PARTIAL TRANSLATION IN A MULTILINGUAL INFORMATION SYSTEM

P. Canisius
Director and Professor
Bundesanstalt für Strassenwesen, Cologne

Abstract

In the Bundesanstalt für Strassenwesen (Federal Institutions for Roads) in Cologne, whose computer centre is at the same time the computer centre of the Federal Ministry of Transport, in past years a series of documentation systems has been built up of which the largest, the International Road Research Documentation (IRRD) of the OECD, is a multilingual one. The IRRD pool, with German, English and French as working languages, today comprises some 90.000 documents and grows annually by approx. 12.000 units. The system, which has been computerized since 1972, suggested the idea to help overcome the language barrier for the user by an automatic partial translation of the abstracts and descriptors of the documents.

With the aid of a "dictionary" a numerical code for the later partial translation is inserted behind every word listed in the "dictionary" in a preparatory run. The translation is then written with the aid of and in the place of this code. The procedure, which can be performed in stages, results through its restriction to technical terms supplemented by a few frequently used substantives and verbs in greater readability of the document for the user, facilitates a later full translation if required and can easily be adapted to new developments in the field of automatic full translation.

1. International Road Research Documentation (IRRD)

1.1 Development of the system

In 1967 the member-countries of the OECD decided to build up an international documentation system as part of the then programme for road research. German, English and French were selected as working languages. The name of the system in these three languages is as follows:

- Internationale Dokumentation Strasse (IDS)
- International Road Research Documentation (IRRD)
- Documentation Internationale de Recherche Routière (DIRR)

The system covers information on every aspect of roads, in particular roadbuilding technology, traffic engineering, accident research and overlapping traffic problems. Today more than 25 institutions from 16 countries are cooperating in the discovery of data. The editing is done in three language centres:

- Bundesanstalt für Strassenwesen, Cologne
- Transport and Road Research Laboratory, Crowthorne, Berks.
- Laboratoire Central des Ponts et Chaussées, Paris

in which the documents have since 1972 also been written on communications format tape (cft). The three "language tapes" are then combined on a master tape, which is sent to all members as monthly IRRD cft. The ultimate users feed the cft into the most varied information retrieval systems. Users without a computer receive hard copies of the data for further processing by hand, e.g. in visual punched card systems. The total pool of the IRRD today comprises some 90.000 documents and grows annually

by approx. 12.000 units.

1.2 Trilingual thesaurus and documentation build-up

Right from the start methods had to be sought in this international system for over-coming the language barrier. In the IRRD a multilingual approach was consistently pursued. In this way at least a certain linguistic proximity of the member-countries working on a language coordination centre was retained.

As a consequence of this system philosophy a trilingual thesaurus was developed which - as a fourth language, as it were - also has a numerical coding of all descriptors. Today 44 subjects have been allocated some 2000 descriptors which are available to indexers and searchers both in an alphabetical list in each of the three languages and via arrow diagrams.

A special thesaurus committee is currently engaged on improving and adapting the thesaurus in all three languages, so as to transform inconsistency of meaning into consistency or at least similarity of meaning and to open up the steady further development of the technical language of traffic for the indexers and searchers of the IRRD system.

In addition to an identification part, the documents in the IRRD pool contain in particular the bibliographical information on the research work or publication covered plus an abstract and descriptors. In general the documents in the pool stay in the language in which they were fed in. However, every centre is free to make translations which can then also become a part of the official pool with a special identification. In particular, in addition to translations into English, Spanish partial translations have recently become a regular

feature.

2. Initial considerations on the automated conquest of language barriers

In view of these partial activities and in the light of the undeniable difficulties both for searching or the searcher and for the later reader of the documents, regarding whom basically good knowledge of all three working languages was assumed, the question also arose for the IRRD pool of overcoming the language barriers by automatic translations.

However, in the German coordination centre difficulties do not occur in searching. The information retrieval system used in Cologne, GOLEM, allowed even in its first version of combining the synonyms in all three languages with the numerical coding, so that on-line search operations performed on the screen could if need be always be done in the German language. However, precisely this linking of synonyms and the technology behind it gave the first indications for a step in the direction of translation aids for the later reader of the documents.

Close observation of the effort involved in automated full translation clearly showed that the variability, the semantic directness, the vagueness, the ambiguity and the dependence on context of the German language in particular would for a long time yet stand in the way of this technique. On the other hand, research work on terminology of technical language demonstrated that in demarcated areas a satisfactorily unambiguous allocation of single words and also whole sentences for the purpose of automatic translation would be possible.

In this state of knowledge a retrospective look at the thesaurus already available in three languages with its

numerical allocation showed the way to a pragmatic if not yet aesthetically satisfying interim solution: a computerized partial translation.

3. The computerized partial translation

3.1 Basic ideas and working method

The possibilities existing so far for automatic word and sentence recognition can be used as follows for a computerized partial translation:

1. First a "dictionary" of relevant words in a basic form (substantives in the singular, verbs in the infinitive etc.) is compiled with translations into as many languages as required and with an additional numerical coding.

2. In a preparatory run for the later partial translation the corresponding numerical code is added to each word that can be derived from the basic form of a word in the dictionary. In the case of output in the original language, this code is suppressed. When the output is in another language (controllable via parameters) code and single words are overwritten by the corresponding basic form from the dictionary.

This method makes it possible, after input into the computer and a single translation run, to print out the document including the descriptors with interspersed aids to translation in all languages contained in the dictionary.

The quality of such a translation depends essentially on the size of the dictionary. The dictionary must above all comprise the thesaurus of the system. If the dictionary also contains verbs and adjectives and a few other

substantives, the method yields good results for the technical and scientific field.

The following are further advantages of a partial translation of this kind:

1. A full translation is considerably facilitated by the interspersed technical words.

2. Users with basic knowledge of the original language of a document can effortlessly understand the text on account of the interspersed words.

3. Progress in the field of automatic translation can be continually used for improvement of the quality of translations.

The advantages cited easily outweigh disadvantages such as the somewhat greater memory and time requirements with extensive texts and the difficulties in compiling and maintaining the dictionary.

3.2 The working method illustrated by an example

In what follows the workings of such a system will be explained by means of a text extract from the IRRD. Fig. 1 contains a dictionary tailored to the text chosen as an example here in German, English and French, including the numerical code.

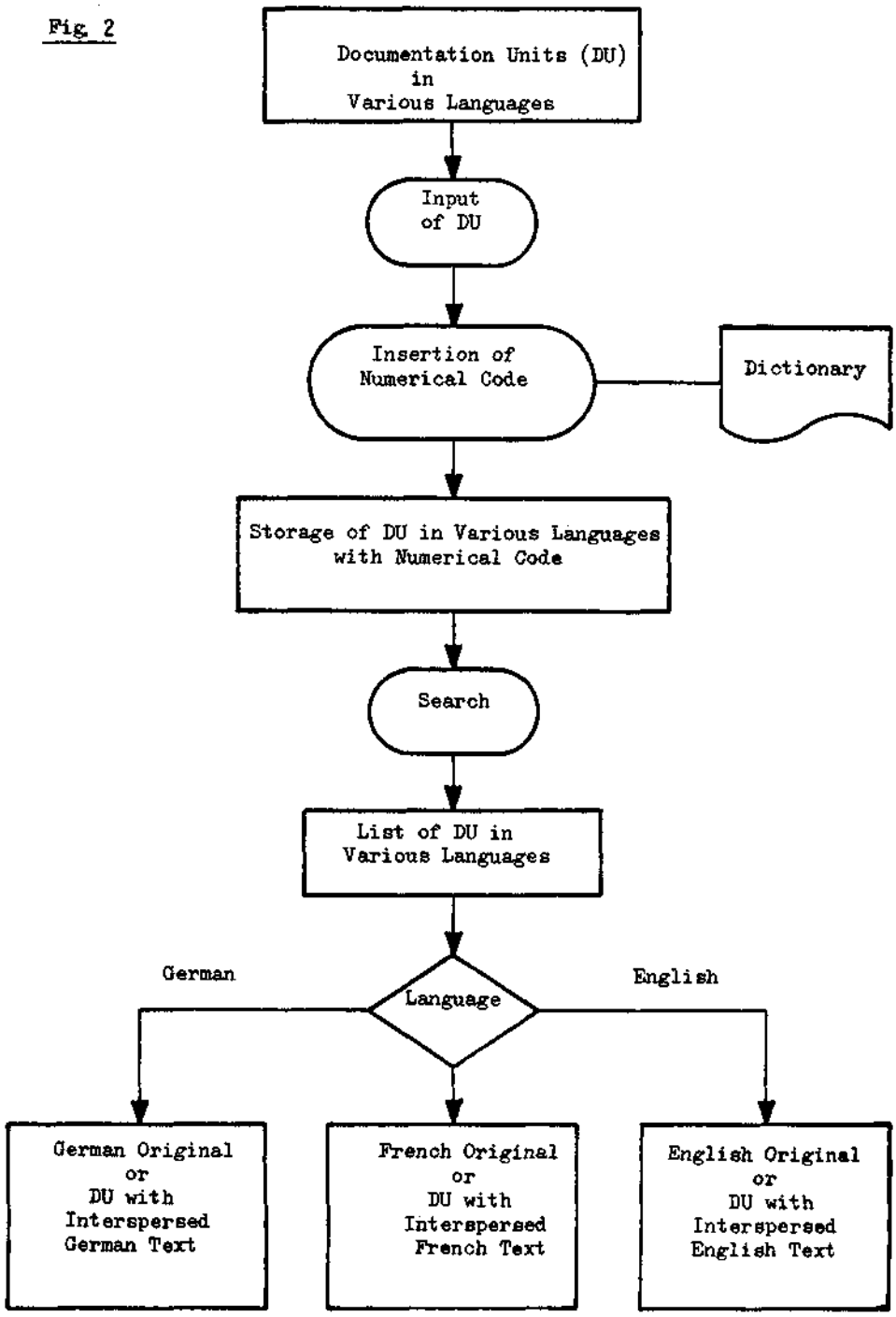
The working method before documentation units (DU) are input up to output of a search can be read off from Fig.2. In the first experimental stage a restriction to the three IRRD working languages is advisable, for which the trilingual technical dictionary part was already available in the form of the trilingual thesaurus. However, via the numerical coding the link-up of as many other languages as required and thus the making of corresponding partial translations are already incorporated in the system.

Fig. 1

DICTIONARY

Numerical code	German	English	French
2853	Problem	problem	problème
4257	Wasserabfluss	water—run—off	ruissellement de 1'eau
1497	Fahrbahnober-fläche	road surface	chaussée
4307	Windrichtung	wind direction	direction du vent
3011	Spurrille	rut	ornière
2899	Querneigung	camber	profil en travers
1701	Länge	length	longueur
1085	Abflussweg	run-off path	trace d'écoulement
3163	Typ	type	type
4010	Verfahren	method	méthode
1787	Messung	measurement	mesure
1653	Griffigkeit	skidding resistance	glissance
2975	Rauheit	roughness	rugosité
1572	Grenzwert	limit	valeur limite
1241	Daten	data	données
3255	Unfallsta-tistik	statistics of accidents	statistique des accidents
1191	beschreiben	describe	décrire
1433	diskutieren	discuss	discuter
1212	bestimmen	determine	déterminer
4132	verschieden	different	différent
3323	uneben	uneven	inégal

Fig. 2


Documentation Units (DU) in Various Languages
Input of DU
Insertion of Numerical Code
Dictionary
Storage of DU in Various Languages with Numerical Code
Search
List of DU in Various Languages
German
Language
English
German Original or DU with Interspersed German Text
French Original or DU with Interspersed French Text
English Original or DU with Interspersed English Text

Below the original text of an IRRD abstract is given as an example (Fig.3)

PROBLEME DES WASSERABFLUSSES VON FAHRBAHNOBERFLAECHEN, WIE WINDRICHTUNG, UNEBENHEITEN UND SPURRILLEN, QUERNEIGUNGEN DER FAHRBAHN UND LAENGE DES ABFLUSSWEGES, WERDEN DISKUTIERT. SIEBEN VERSCHIEDENE TYPEN VON FAHRBAHNOBERFLAECHEN MIT VERSCHIEDENEN VERFAHREN ZUR MESSUNG IHRER GRIFFIGKEIT UND RAUHHEIT WERDEN BESCHRIEBEN. PROBLEME DER BESTIMMUNG VON GRENZWERTEN DER FAHRBAHNGRIFFIGKEIT WERDEN DISKUTIERT, WOZU DATEN DER UNFALLSTATISTIK HERANGEZOGEN WERDEN.

Fig.3

In the following text (Fig.4) the numerical coding has been inserted:

PROBLEME(2853) DES WASSERABFLUSSES(4257) VON FAHRBAHNOBERFLAECHEN(1497), WIE WlNDRICHTUNG(4307), UNEBENHEITEN(3323) UND SPURRILLEN(3011), QUERNEIGUNGEN(2899) DER FAHRBAHN UND LAENGE{1701) DES ABFLUSSWEGES(1085), WERDEN DISKUTIERT(1433). SIEBEN VERSCHIEDENE(4132) TYPEN(3163) VON FAHRBAHNOBERFLAECHEN(1497) MIT VERSCHIEDENEN(4132) VERFAHREN(4010) ZUR MESSUNG(1787) IHRER GRIFFIGKEIT(1653) UND RAUHHEIT(2975) WERDEN BESCHRIEBEN(1191). PROBLEME(2853) DER BESTIMMUNG(1212) VON GRENZWERTEN(1572) DER GRIFFIGKEIT(1653) DER FAHRBAHN WERDEN DISKUTIERT(1433), WOZU DATEN(1241) DER UNFALLSTATISTIK(3255) HERANGEZOGEN WERDEN.

Fig.4

In Fig.5 the codes are replaced by English translations:

PROBLEM DES WATER-RUN-OFF VON ROAD SURFACE, WIE
WIND DIRECTION, UNEVEN UND RUT, CAMBER DER FAHRBAHN
UND LENGTH DES RUN-OFF PATH, WERDEN DISCUSS.
SIEBEN DIFFERENT TYPE VON ROAD SURFACE MIT DIFFERENT
METHOD ZUR MESUREMENT IHRER SKIDDING RESISTANCE UND
ROUGHNESS WERDEN DESCRIBE. PROBLEM DER DETERMINE VON
LIMIT DER SKIDDING RESISTANCE DER FAHRBAHN WERDEN
DISCUSS, WOZU DATA DER STATISTICS OF ACCIDENTS HERAN-
GEZOGEN WERDEN.

Fig. 5

Fig.6 gives the same partial translation into French:

PROBLEME DES RUISSELLEMENT DE L'EAU VON CHAUSSEE,
WIE DIRECTION DU VENT, INEGALE UND ORNIERE,PROFIL
EN TRAVERS DER FAHRBAHN UND LONGUEUR DES TRACE
D'ECOULEMENT, WERDEN DISCUTER. SIEBEN DIFFERENT
TYPE VON CHAUSSEE MIT DIFFERENT METHODE ZUR MESURE
IHRER GLISSANCE UND RUGOSITE WERDEN DECRIRE.
PROBLEME DER DETERMINER VON VALEUR LIMITE DER
GLISSANCE DER FAHRBAHN WERDEN DISCUTER, WOZU
DONNEES DER STATISTIQUE DES ACCIDENTS HERANGE-
ZOGEN WERDEN.

Fig.6

By a corresponding control instruction an output confined to the translated individual terms can be obtained. A look at such an output (Fig.7) shows that in this way too preliminary information can be supplied on the contents of the document.

```
PROBLEM    WATER-RUN-OFF    ROAD SURFACE,
WIND DIRECTION, UNEVEN    RUT, CAMBER
    LENGTH    RUN-OFF PATH,       DISCUSS.
       DIFFERENT TYPE    ROAD SURFACE    DIFFERENT
METHOD    MESUREMENT      SKIDDING RESISTANCE
ROUGHNESS       DESCRIBE. PROBLEM    DETERMINE
LIMIT    SKIDDING RESISTANCE
DISCUSS,     DATA    STATISTICS OF ACCIDENTS
```

Fig.7

3.3 Staged implementation

The first step in the direction of such a system with respect to the IRRD can be taken without involving much work by using the thesaurus with numerical code as dictionary. This procedure makes it possible to print out the descriptor part in each of the three permitted languages, by which the user can gain an approximate idea of the contents of the document. In a second step the thesaurus must be supplemented by frequently required substantives and verbs not specific to the subject. As numerous works have already appeared on this theme, it is simply a matter of transferring the results found there so as to supplement the dictionary.