Hermann-Josef Zingel
Zentralstelle für Textildokumentation und -information, ZTDI, Düsseldorf

Experiences with TITUS II*

Zingel, H.J.: Experiences with TITUS II.
In: Intern. Classificat. 5 (1978) No.l, p. 33-37

Description of the international cooperative documentation system called TITUS (Textile Information Treatment Users' Service) in its previous and present form (TITUS II). It uses a special linguistic way of automatic translation of abstracts and index terms (with a controlled vocabulary and a controlled syntax) in order to supply users of the English, French, German or Spanish language with abstracts in their native language from inputs in one of the other languages. I.C.

1. Cooperation in international textile documentation

The acronym TITUS stands not only for a translation system but also for a network for international cooperation in the field of textile documentation. The acronym TITUS is an abbreviation of "Textile Information Treatment Users' Service". In Germany we call it: "Textil-Informationstechnik und -service".

The TITUS II translation system comprises so far the four languages English, French, German and Spanish. In 1970 a system of international cooperation in the field of textile documentation was set up which encompasses nearly all Western European countries and has since been joined by the United States as well. The computer center of this international network is located in France, in Boulogne-sur-Seine near Paris. To the computer installed there, all cooperating partners have access in order to provide the input and use the output. A division of labor has been instituted whereby the partners feed in abstracts for the data base, with some of them, e.g. the German "Zentralstelle für Textildokumentation und -information" (ZTDI = Central Agency for Textile Documentation and Information) in Düsseldorf employing to this end the teleprocessing dialog mode. For the input of these abstracts the feeder can employ any one of the four working languages, according to his preference. Similarly, the system permits the output to be printed out in any one of the four working languages, depending on the user's preference.

* We gratefully acknowledge the permission of the Deutsche Gesellschaft für Dokumentation e.V. to print this revised English version of a paper presented at the "Deutscher Dokumentartag 1977" in Saarbrücken. For the previous version see the Proceedings volume of this conference.

At this point we will dwell briefly on some general questions of the international textile documentation scenery which illustrate the connections existing with the translation system itself. The steady expansion of knowledge finds expression in an immense output of publications, data collections and documents which are often difficult to procure and which may be printed in any of up to a hundred languages. It is especially in the scientific and engineering fields that the flood of information is steadily increasing. Here the amount of information available is truly snowballing, increasing as it does at growth rates which in areas of particularly stormy development produce a doubling of the information quantities every three years. There is no area left in which this rate lies below 8 %, a percentage meaning a doubling within a period of nine years.

The textile industry, too, is a field abounding in information. Some 800 technical journals, 500 books, 500 reports and 10.000 patent specifications appear every year. Individual publications number up to some 100.000 each year. Even if duplicate and multiple publications, especially in journals, are weeded out, there still remains a balance of some 30.000 publications each year for our industry alone.

In addition, the main producers of scientific and technical information and documentation have come more and more to the conclusion that, in this field of activity, complete national self-sufficiency does not constitute a realistic alternative to prevailing policies. Everywhere it is becoming quite clear that the many and manifold information problems cannot be solved by the individual nations singlehandedly, if only for lack of sufficient resources but most of all for financial reasons. Scientific and technical information is, by its very nature, dependent on cooperation among different countries. Things have already progressed to the point where there is hardly any noteworthy information activity going on which is not conducted on a bilateral, multilateral or international basis. And the pertinent activities taking place on the international level are of a truly staggering complexity.

In the Federal Republic of Germany, all activities in the field of textile documentation have, on the initiative of the "Wirtschaftsverband Gesamttextil" (= Overall Economic Association for the Textile Industry), been concentrated into a single organization so as to permit efficient cooperation with international partners. The mutual interests shared with other countries, especially with France and its "Institut Textile de France" (ITF), complemented each other in an excellent way. It was the ITF which had conducted the initial tests on computer documentation and translation. These tests were also conducted in the interest of the afore-mentioned ZTDI agency, a branch of the "Verein Deutscher Ingenieure" (= Association of German Engineers).

Subsequently we jointly established an international network which started out by instituting an international division of labor for the input of abstracts (Participating countries: France, Spain, Italy, Belgium and (West) Germany). The input into the textile data base was furnished through international collaboration, with each participating country taking care of a certain portion usually determined by the language used. Within the limits of their ability, the individual countries operate as

centers for the output not only for themselves but also for neighboring and other countries. The countries predominantly involved are West Germany and France. The ITF, at its headquarters near Paris, provides the computer. Via a terminal and a data line, our German documentation center in Düsseldorf has an interactive system for the purpose of on-line storage of abstracts and on-line retrieval of the desired information from that collection. Since the beginning of the teamwork, approximately 90.000 documents have been stored. Every year, some twenty to twenty-five thousand new ones are added. Since the early days of international cooperation in textile documentation, and after the time of putting the TITUS computer system into operation, the system has gone through various stages of development. Each one of these stages represents a working step in the development plan of the overall project.

2. First phase: abstracts in one language only: TITUS I

In the initial "TITUS I" stage from 1970 to 1973, a multilingual thesaurus with descriptors in the five languages German, English, French, Italian and Spanish was applied. Using this multilingual thesaurus, the indexers stationed at the centers of the international network could index the documents in the language of their choice and prepare conventional abstracts, likewise in one of the five languages. In the output, however, only the descriptors could be translated into one of the languages. The abstract was only available in the language of the TITUS partner who had fed the information into the central collection in the first place, or in French.

Experience with our users confirmed that this matter of foreign language abstracts was a considerable obstacle to the exchange of ideas and the transfer of information. The users' satisfaction with the completeness of the information collected was more and more being marred by the fact that this information was largely presented to them in a language differing from their native one.

Unfortunately, in the area of textiles we are not dealing with an English-language-oriented field where scientists or practitioners might readily agree on using English as a common, neutral working language. It is somewhat risky, anyway, to believe that one single language, e.g. English, can act as the universal carrier of knowledge in a technical specialty. As we know, nothing but complete mastery of the language concerned will do if one is to grasp, for instance, the full implications of a legal text or the subtleties of an invention hidden in a patent.

Only those readers whose native language is identical with the language of the document will be able to grasp the essentials of the document in all their implications.

One more imperfection of an internal documentation nature turned up during this first phase of TITUS. All partners to the TITUS project employed indexers working externally on document analysis: out of a total of 77 individuals, 11 in West Germany alone. In spite of the fact that all these indexers were highly qualified specialists in their specific fields, the effectiveness of retrieving the documents indexed by these specialists turned out to be inadequate, even though uniform indexing rules were used. There always remained some non-homogeneity in the descriptor selection and a remnant of diverging indexing depth. However this working procedure, highly classical as it was, made it possible to establish and consolidate the basic network and also to gather the necessary experience for the development of the next stage, TITUS II (see Fig. 1).

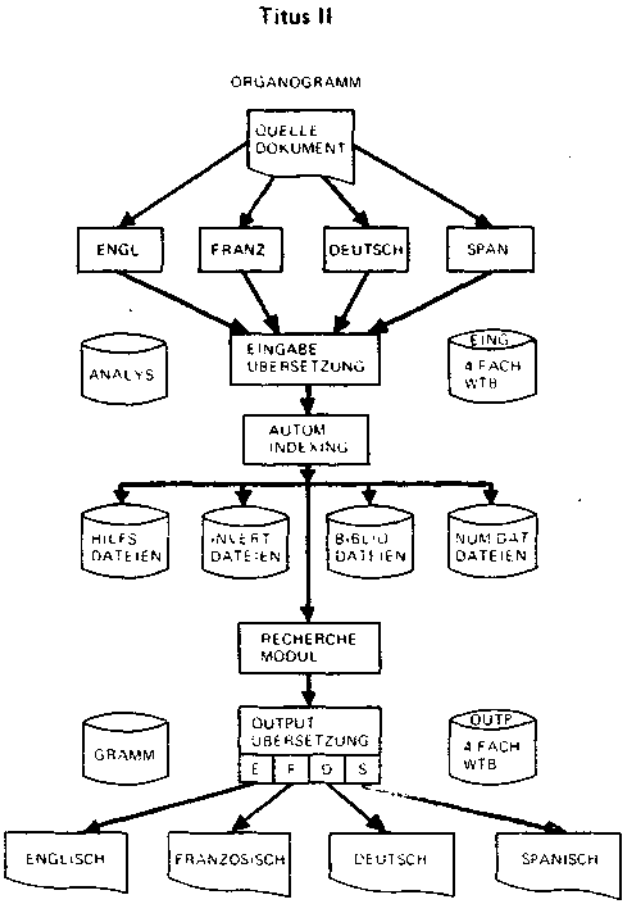

Fig. 1: Organogramm of TITUS II

3. Abstracts in native language of users: TITUS II

It was our agreed objective to improve the analyzability of documents by means of a multilingual translation system, and simultaneously to supply the users of the information system with the desired information in their native language. It was, moreover, very desirable to achieve uniformity of document analysis so as to improve the accessibility of the documents to be searched.

Political reasons, too, were involved in tackling and solving the problem of automatic translation. Truly international cooperation in the information and documentation field can only be ensured if there is neither discrimination against nor prevalence of any single language. International information transfer also depends upon international feedback. Therefore, automatic translation also ensures a wider dissemination of the literature of the home country. In the home country itself, furthermore, information from abroad is made accessible, through the translation, to a larger scientific community, especially in the engineering field.

In all languages, there is a limited number of syntactical rules which permit the formation of an unending number of phrases by using a very great but limited number of lexems. Knowledge of the vocabulary, the grammar and the internal phrase structure is assumed

here. So far, linguists have only partly succeeded in explaining this structure.

To this moment there exists no complete, self-contained theoretical basis for truly automatic translation. But there is Chomsky's theory about language, a theory which was very important for the elaboration of the TITUS translation system: it permitted an attempt to reduce the number of syntactical rules in order to develop a universal linguistic base model for the mechanical translation of information into several languages. This linguistic model must be a very simple one, because in the field of scientific and technical documentation the objective is to transfer information rather than to translate literary texts with all their niceties and spoonerisms.

Finally, in 1973, the efforts to develop a linguistic system for the translation of abstracts were crowned with success. Ever since the 50's, all research conducted to permit the mechanical translation of texts was geared to providing means for translating all free-written texts. TITUS employed the opposite approach, an approach providing that normal, free-written technical texts will be rescripted using words, rules and structures which the computer knows. The language which the resulting rescripted text employs is called by us "Canonical Documentation Language" (CDL).

4. Our "Canonical Documentation Language"

The special feature of our Canonical Documentation Language is the use of a controlled syntax which has to be coded for document analysis, together with the controlled textile vocabulary which is available in four languages, namely German, English, French and Spanish. Moreover, words of ordinary language are needed in order to express associations of thought in the technical text. In a later transition stage towards a further development of the TITUS system, the codification of syntactical relations and other language parameters should no longer be necessary.

The canonical documentation language aims at nothing else but the possibility to clearly arrange a lingual expression according to fixed grammatical rules. For this arrangement a number of phrase models is on hand. This canonical phrase can consist of up to five propositions. The syntactical relations between these propositions are established by actants. Each proposition consists of lexical units from the quadrilingual thesaurus. Within the proposition, the lexical units are likewise in syntactical relation. In this manner, skeletons will be built whose unoccupied places will be filled in by lexical units of the technical or ordinary language. The phrase-skeleton will be decoded by the computer.

Figure 2 shows the basic model of a sentence for automatic translation. Each arrow means a syntagmatic group, i.e. a proposition or part of a phrase. It can consist of up to four lexical units from the thesaurus. These groups are brought into syntactical relation by the actants. Our actants are prepositions or prepositional phrases which effect the flexions of the sentence or the proposition. The actants, which will be coded, are placed before the phrase. Group A and group B are always in genitive case relation, because group B is used as a subject complement. At present, about 80 actants are allowed.

The syntactical relation between the individual lexical units within a proposition will be coded in the proposition itself. There are 10 such relators, as we call them. Numerous combination possibilities are provided by exchanging actants and relators. As a result, there is an extremely wide range of different phrase constructions.

Special difficulties were presented, however, by the German language. For one thing: words are declined in German, and all endings of declinations had to be put into the machine. We learned that there are relatively few rules for the declination endings. The formation of composite words follows no uniform rules. In German, in addition, it has to be taken into consideration that there are prepositions which rule the accusative or dative case, depending on whether the phrase reflects a

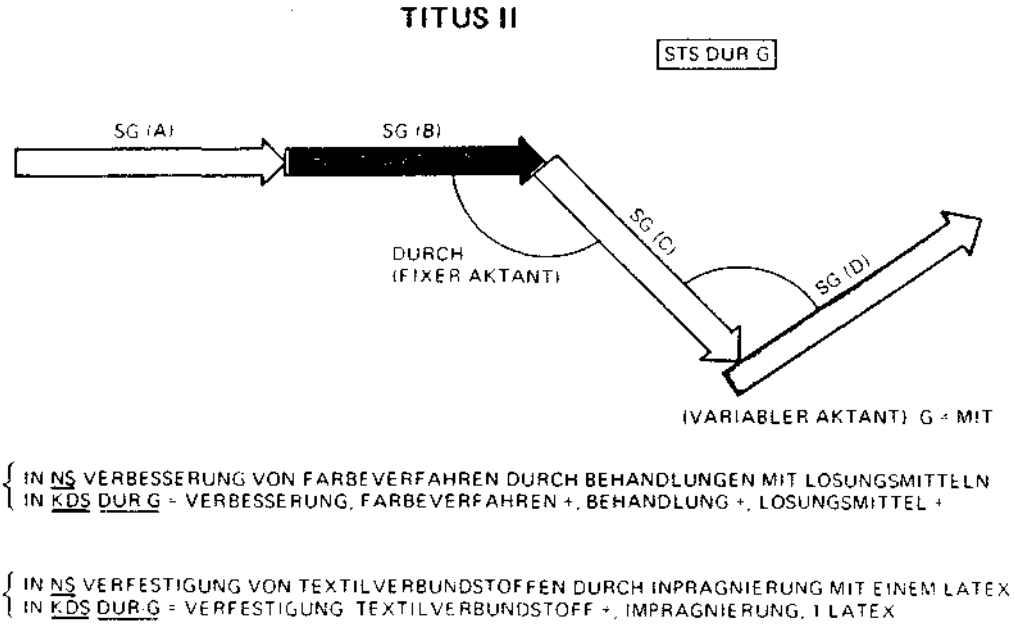

Fig. 2: Basic model of a sentence for automatic translation in TITUS II

movement or a non-movement. This will not be expressed in like manner in the other languages of the system.

5. A multilingual vocabulary with different kinds of lexical units

Now I would like to say something about the vocabulary we use in our information system. Our vocabulary consists of different kinds of lexical units. Its bulk is made up of the technical terms related to the textile industry. These terms are the descriptors. A further part is formed by the tool words without documentary value but necessary for the formation of correct and understandable sentences. Then there are adjectives, both qualitative ones such as black, white, red, etc., and attributives ones such as some, few, several. And furthermore there are the verbs, but these are restricted to transitive ones so as to avoid the difficulties presented by the intransitive verbs in German. Last but not least the vocabulary includes the prepositions and prepositional expressions, while it also contains all forms of the definite and indefinite articles. All in all, and excluding synonyms, the vocabulary comprises some 11.000 words.

All descriptors, tool words, adjectives and verbs are lexical units. Each lexical unit is a written representation of a thought, or rather: of an exact concept, and it may be represented by one or by several words. It provides as complete as possible a translation into the other languages of the concept which it represents, rather than of the words themselves of which it consists.

These lexical units must not be polysemic. A definition adds a precise meaning in the case of a possible polysemy. The actants, i.e. mainly the prepositions, have an equivalent in each language, but this cannot always be regarded as having absolute validity.

This translation model allows for automatic translation without producing telegram-style translations. A transformation grammar, developed for every language as a so-called language-specific syntax filter, provides already today each language concerned with a syntax of its own for the information printouts in that language. All stored information is transformed into a swivel language in binary form. From this, the output will be recorded and translated into the target language. Each document will be indexed automatically based on the text of the abstract. The computer automatically identifies the descriptors and builds up the corresponding files.

Figure 3 shows the characteristic steps of TITUS II as performed by the machine and the documentalist.

6. Preparation of abstracts

Now what is our experience with this system? Outsiders, as well as some documentation specialists, suspects sometimes that the preparation of text as we do it right now is not transferable to other fields, and moreover that this can only be done by people who do it every day. Furthermore it is felt that the preparation of these texts takes too much time, since the indexer is only permitted to use words known to the computer, i.e. appearing in the vocabulary, thus being squeezed into so rigid a language system that his abstracting job, which is not an easy one anyway, ceases to give him any pleasure at all. I must admit that this actually did worry us while the system was being developed. We were in particular concerned that the indexers in charge of preparing the abstracts for us would quit. This, however, did not happen. Quite on the contrary! The situation here is rather like playing chess. Fixed rules have to be followed, but nevertheless the game can be great fun.

In the TITUS network our German center and the French one are coupled with each other with functional centralization being maintained through a private data line. Both partners operate the computer in the same fashion, except that we do it in German and ITF in French. This pertains to input as well as to output. The online connection to the computer center continuously updates, of course, the central store by feeding into it the information of the partner concerned. The user profits by this. The value of all document abstracts is the same, both as regards their contents and their formal description. In principle there are no differences whatsoever. This is assured in particular by the fact that the very system for the linguistic abstracting and transformation was designed in such a way that, due to the formalities to be observed in the preparation of abstracts, hardly any differences can be produced from one case to another, provided the contents of publications are correctly interpreted and employed in the abstract.

7. Outlook and conclusion

Due to the permanent on-line dialog with our data base we are in a position to answer incoming requests immediately. The linguistic presentation of the abstracts is considered to be between satisfactory and good and this for the whole range of the different information services.

Fig. 4 shows the characteristic steps performed by the machine and the documentalist, as well as the merits and shortcommings of an expected new linguistic TITUS model, a system closer to free-text translation but still imposing restrictions and rales, for the rewriting of a text.

Now, to conclude:

— TITUS II is a system for the translation of documentary texts with a controlled multilingual vocabulary and a controlled syntax.
— In contrast with conventional documentation systems, there is no division here between the indexing process and the writing of an abstract. In TITUS, both form one single, integrated processing step, and this not for just one or two but for no fewer than four languages.
— The significant terms are descriptors for building up the search files.
— The grammatical structures employed represent not all possible structures occurring in ordinary language but only a part, and these structures are clearly defined.
— The translation achieved will be near-perfect, without any danger of technical misunderstandings.
— Storage in the computer is in binary form. So are the lexical units and the structures of the phrases.
— The original title of an article will not be automati-

cally translated unless it is rescripted, which is entirely possible.

In comparison with the concepts of free-text translation systems, TITUS has the following advantages for information systems:

_ A limited and readily overseeable vocabulary which due to its brevity is easier and cheaper to store than normal vocabularies.

- A controlled grammar, reducing the required storage capacity even further. Furthermore the phrases are simpler, shorter and cheaper.

The writing of an abstract in CDL instead of a free-text abstract means hardly any loss of time, especially if the freetext abstract has to be indexed in addition anyway.

All other things being equal, the comprehension and abstracting of the document essentials by a documentalist in his own native language presents less difficulty and produces better results.

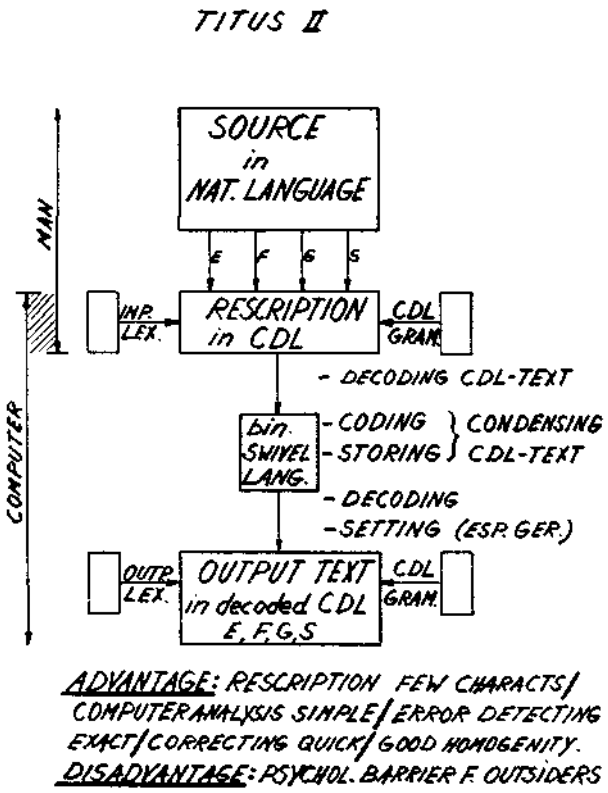

Fig. 3: The characteristic steps of TITUS II

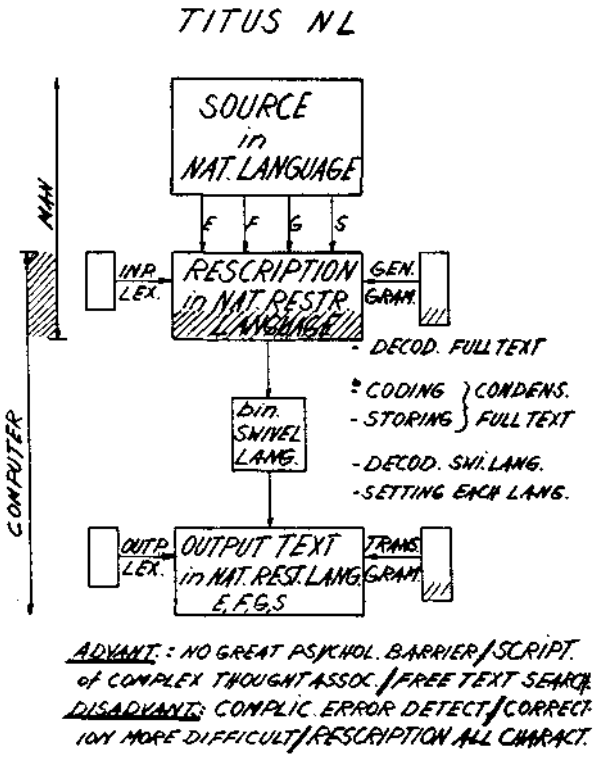

Fig. 4: A possible new linguistic TITUS model (TITUS NL)