[From: *Translating and the Computer 14.* Papers presented at a conference… 10-11 November 1992 (London: Aslib 1992)]

USING METAL - THE APPLICATION OF A MACHINE TRANSLATION SYSTEM IN AN R & D ENVIRONMENT.

Alain Paillet

The introduction of an MT-System as a tool supporting the human translator in a corporate translation Department has repercussions on day-to-day business with respect to acquiring the skills to operate the system and modifying the organization of work. There are a number of prerequisites to be met before acquiring a System like METAL. The report covers all phases from the analysis of text material with regard to its suitability for processing by MT to the subsequent text production with METAL.

INTRODUCTION

My topic is the use of METAL in an R&D environment and my purpose to deliver a user report. I have divided it into four parts. I will first describe the corporate environment at Boehringer Ingelheim, derive from this the need for linguistic communication and explain how we meet the requirements.

I will then formulate 4 prerequisites for the use of METAL and show how we have implemented the system at Boehringer Ingelheim.

Finally I will deal critically with the problems encountered so far with METAL and with the important aspects of that artificial intelligence system.

Corporate Environment and Need for Linguistic Communication

Boehringer Ingelheim is a family-owned pharmaceutical company with subsidiaries in all five continents. The Headquarters are based in Germany in Ingelheim am Rhein between Mainz and Bingen.

Here are the main figures:

FIGURE 1

The level of R & D costs shows how important this field is to the company.

The main indications for which Boehringer Ingelheim produces pharmaceuticals are diseases of the cardiovascular system and the respiratory tract such as asthma and bronchitis. Over the past years activities have also focussed on the central nervous system with products for a form of senile dementia, better known as Alzheimer's disease.

In the cardiovascular field, Boehringer markets a product manufactured by genetic engineering technology. This is a fibrinolytic agent that enhances the recanalization of coronary vessels after infarction.

Marketing authorization for a pharmaceutical product is always preceded by the submission of a registration file to the competent authorities. In this file all development phases of the product must be documented, from the analytical phase where the molecule is isolated through production and quality control to clinical trials in healthy volunteers and patients. We are talking about thousands of pages. But that file is only valid for a certain dosage form (e.g. tablets or syrup) and a specific indication. If you want to apply for marketing authorization for ampoules in the same indication, you need another file. As our products are registered worldwide, you can imagine the volume of documentation involved.

As the Boehringer Ingelheim Corporation has five research plants worldwide and four production plants in Europe you will understand the importance of linguistic communication.

In 1991 the volume of translation amounted to 34 000 pages. The volume in 1992 will be in the same order.

Here are the principal language combinations

FIGURE 2

How do we meet these requirements?

We have a translation capacity of 13 man/years spread over three sites: in Ingelheim, at BIL UK in Bracknell and at BI Spain in Barcelona.

Carefully selected freelancers help us to cope with volume peaks.

In 1991 overall costs for translation by corporate translation department and freelancers amounted to 4.4 million Deutschmarks, that is roughly 1.5 million pounds.

Given this translation volume and these costs, we needed a tool beyond mere word processing and terminology management which are quite standard today. The tool we looked for should enable us to translate large volumes of text with fairly straight-forward syntactic structure.

We therefore investigated METAL, decided to test the system for the language combination German-Spanish for five months and by mid 1991 we finally acquired the system with two language combinations, German-English and German-Spanish. METAL stands for "Machine Translation and Analysis of Natural Language".

I will not go into details as to why we opted for METAL and not for another system. Any potential user should decide that for himself. He may for instance ask the recently founded European Association for Machine Translation for its support. Being a central service group we know what kind of texts are to be translated and the volume involved. We also know the morphosyntactic characteristics of these texts.

This is a very important, perhaps the most important prerequisite for use of an MT-system. But there are more.

PREREQUISITES FOR USE OF AN MT-SYSTEM

I should like to point out that these prerequisites are not specific to METAL. They would probably apply to any MT-System.

FIGURE 3

Firstly:

You need suitable texts in sufficient quantity.

The texts are suitable if they are of a declarative character and a simple syntactic structure. By simple syntactic structure I mean a main clause with a limited number of subordinate verbal phrases. Technical documents, instructions for use, specifications, technical reports fit into this text category.

By "sufficient quantity" I mean that the regular volume should be large enough to allow return on investment within a limited number of years.

Secondly:

Make sure that you have as much of the text volume as possible in a machine readable form. There is no point in losing on the roundabouts what you gain on the swings by keying the texts into the machine. In this case human translation is cheaper.

Use a scanner and an OCR-software with learning capacity. There are a number of good ones on the market in a price range between 500 and 1000 pounds.

Thirdly:

Make sure that your MT-system can be properly interfaced to the surrounding world. Isolating the system would defeat the communication aspect of translation. What sounds easy in technical descriptions often turns out to be very tricky in real world conditions.
At Boehringer Ingelheim, METAL is hooked up in a Novell-LAN.

Fourthly:

You need to have well motivated staff with an interest in machine translation. Staff that are prepared to acquire the appropriate knowledge for proper system operation. There is a lot to learn that is not normally on the syllabus of a translation course at university. You just cannot make full use of the system if you do not know how it performs word and sentence analysis because the user can interfere at that very level.

If you first have to convince your staff to use an MT-system in your department, then you do not have an easy job as in the initial phase you have to invest a lot of time and effort with no immediately apparent return. Your work situation even deteriorates as you are using human translation capacity for a system that is not yet productive. So it is very easy to cast an unfavourable light on the MT-system.

I was lucky not to have had such acceptance problems.

Now I want to explain how we at Boehringer Ingelheim have put into practice the above mentioned four prerequisites in order to make productive use of the system.

AGAIN FIGURE 3

Prerequisite 1

The texts that we process with METAL are, according to the scheme of text typology of Egon Werlich*), so called "non fictional texts" which means that in order to understand them both the author and the addressee must share the same reality model and both have comparable technical knowledge. As an intermediate between the German author and the English or Spanish speaking addressee the translator on the one hand and METAL on the other must also make this technical knowledge available. METAL does it at lexical level in that it provides for terminological consistency, and the system operator in that he uses his knowledge of the subject when post-editing METAL output.

Here is an extract from a stability report in German.

FIGURE 4,

This type of text is characterized by an obvious invariance in syntactic structure between the German original and the English and Spanish translations. A comparison of the parameters "number of words per sentence" and "number of finite verbs per sentence" in all three languages results in the parameter "length of clause" that is represented in the following diagram:

FIGURE 5

So from the point of view of syntactical analysis this kind of text is bound to turn out quite well.

We have an annual total of over 10 000 pages of text of this type for translation into English and Spanish. Although we only had one METAL work station available, we have translated several hundred pages of stability reports into Spanish and preclinical documents into English with METAL. This was possible as we had encoded the appropriate terminology into the system. I will come back to coding later on.

But the breakthrough only came a couple of months ago once we had eventually succeeded in increasing the number of work stations to three.

My group is a profit center and the economic basis for the investment in METAL of roughly 120 000 Pounds is to top up our productivity by 1 500 to 2 000 pages a year of METAL translations, thus allowing at least to earn back the acquisition costs of the hard- and software within a limited number of years. Our ultimate aim of course is to reach a level of productivity enabling us us to cut the price per line.

We shall not be able to reach 1500 pages in 1992 owing to technical problems.

Prerequisite 2

How do we get the texts for METAL into the machine? Reality is very frustrating. Despite all the talk about the paperless office, the reality in our company is still 90% paper, although one does occasionally receive recent texts on magnetic data support. Having METAL hooked up in a Novell local area network, I may be somewhat optimistic for the future.

But nevertheless we use a scanner with an OCR software called OMNIPAGE PROFESSIONAL because it is often less frustrating to produce your own ASCIIS than to wait for the customer to provide you with a machine readable file.

We can scan an A4 page of laser printout quality, including checking, in roughly 1 minute. Experience shows that it is precisely at those interfaces where optimization is required; otherwise you run the risk of losing much of the time saved by translating with the machine.

This brings me to prerequisite number 3 about interfacing.

Prerequisite 3

This is our text production line.

FIGURE 6

This is our METAL configuration. Sietec Consulting have taken over the marketing of METAL from SNI. They have created a new system platform in which METAL uses Sun workstations. The interface aspect might be a bit different but all users I know also have our configuration.

The input and output interfaces are the most important, from the point of view of both the hard and software.

HARDWARE:

We did not succeed in transferring files from a DOS-PC to the SINIX PC using a board which would have allowed us to avoid using SNI monitors. The board suggested by SNI themselves failed to transfer certain foreign language characters; these did not appear on the screen. So we had to go back to SNI monitors. At this point I should like to mention that the support received from SNI was not exactly up to the mark.

SOFTWARE:

Since there is no Word-format converter we can only process ASCII files. This means that the format is always lost. This is most annoying because it inflates post-editing times and has nothing to do with translation. According to Sietec, the Word-format converter should be available soon.

Given these problems, how do you interest your staff in METAL ?

To put it bluntly, you need a couple of computer enthusiasts. Learning to operate the system requires substantial effort. The basic training to understand the architecture and the functioning of the system takes two and a half days. You also need a 3 day coding course during which you learn basic elements of the METAL phrase structure grammar and how to encode the different word categories and especially verbs. With some practice encoding a verb with all its valency aspects takes about 4 to 5 minutes.

Ideally you should also entrust one of the users with the responsibility of system administration which includes backing up data, adding new users, administration of privileges, error diagnosis, trouble-shooting and installing system updates. For this there is a 3 day training during which you acquire the necessary knowledge in two different operating systems.

We now have three translators trained to use METAL but another four staff are involved at different times in file input and/or output.

DRAWBACKS AND ADVANTAGES OF METAL

Drawbacks

When you first buy the system, it is not productive. You must adapt it to meet your requirements. The system dictionary comprises about 30 000 entries of general vocabulary, some common technical vocabulary and some EDP terms, but your own terminology is obviously missing. This must be encoded by the user. We criticised the lexical coverage of the system with respect to verbs as we found that there were not enough verbs in the basic vocabulary. This has been improved in the last release.

We have had hardly any support from the system developer with respect to interfacing METAL to our Novell-Network. The interfacing of software is still not possible; we can only process ASCII files. There is a dire need for converters to the most common word processors.

Amazingly enough, these are marketing aspects that were totally neglected by SNI.

Ever since Sietec Consulting took over from SNI things seem to have improved a lot.

One important point is certainly the fact that the introduction of METAL in a translation group results in most cases in a reorganization of work. The system is now bound to drain all suitable texts, leaving the human translator with some residual capacity to tackle other texts in different subject areas. So, in the end, what was thought to be a drawback might well turn out to be an advantage.

Advantages

But METAL has other advantages:

The system guarantees almost absolute terminological consistency over hundreds of pages in a subject area where drug safety and quality control require precise statements and information. This is also facilitated by the possibility of establishing preferences as to how the system should retrieve data from the various dictionaries.

Time consuming functions such as preanalysis or translation (roughly 10 pages/hour) can be initiated in the evening and run over night.

You can handle large volumes of text and increase the productivity of the translation department.

Any pre-edited German text for translation into English is very often requested in Spanish at a later date. So it makes sense to keep an archive with pre-edited German texts for possible translation into Spanish.

We have a check list for every text processed with METAL in which we record the time taken for every step from input to output:

FIGURE 7, 7a, 7b

Thus we can monitor the optimization of every step in terms of time reduction.

METAL is a product undergoing constant development. There is a forum in which users and system developers meet twice a year to exchange views and set priorities with respect to development, training and support. These meetings have proved to be very productive.

For instance, Sietec has built up a text corpus - as a test-suite - representative of the texts every user processes with METAL in order to be able to test every new programme update in real life conditions before delivery.

But the most important advantage of METAL cannot be expressed in monetary terms, at least not at the beginning. The greatest advantage lies in the creation of a knowledge data base which increases with time and puts the company in a position to produce product documentation independently of personnel turnover and imponderable circumstances. METAL then becomes an important tool for know-how storage and exchange within the Corporation.

This is the real dimension of METAL.

CONCLUSION

I think that regardless of the industrial sector or the institution in which we work, we cannot simply reject the possibility of using an MT-System and carry on as usual as if such systems did not exist.

Bearing in mind the four prerequisites formulated above, METAL can provide a useful tool with which to relieve the donkey work involved in some translations and improve the productivity of a translation department.

Nevertheless, taking on METAL must be well planned in order to avoid disrupting the working process in an attempt to improve it.

METAL should be considered as a step in a document production process and hence it is important to take a closer look at how original documents are written at a previous stage. Perhaps METAL will help you establish that these documents are not written properly when one considers their purpose, namely to serve as a carrier of clear and concise information. Sentences are too long (any sentence exceeding 30 words must be read twice), they contain grammatical errors, punctuation is not consistent, etc... . You will then discover out that any step taken to improve the quality of these documents also has a direct and positive effect on METAL-output. We have not progressed that far yet, but we are thinking of setting up a few basic rules for authors.

What are the prospects with METAL in the future? Well, there is no cause to be euphoric, we're just cautiously optimistic.

BOEHRINGER INGELHEIM CORPORATION	1991
NET SALES	5,225
Of these: in Germany	25%
abroad	75%
Of these: human Pharmaceuticals	80%
other business activities	20%
RESEARCH AND DEVELOPMENT	816
NUMBER OF EMPLOYEES	24,347

Figure 1

TRANSLATION VOLUME: 34 000 PAGES / YEAR

LANGUAGE COMBINATIONS

DE - EN

DE - SP

DE - FR

EN - SP

EN - FR

SP - EN

IT - EN

60 % into English

Figure 2

1. DECLARATIVE TEXTS IN SUFFICIENT QUANTITY

2. TEXTS MUST BE MACHINE READABLE

3. APPROPRIATE INTERFACES

4. MOTIVATED AND INTERESTED STAFF

Figure 3

BUSCOPAN COMP. SUPPOSITORIEN

GEHALT METAMIZOL NATRIUM: METHODE

UV-spektrophotometrische Bestimmung

Probelösung:

5 Suppositorien werden in einem 500 ml-Meßkolben mit etwa 100 ml Wasser im Wasserbad von ca. 40 °C geschmolzen, anschließend 15 Minuten geschüttelt und mit Wasser zu 500,0 ml verdünnt. Etwa 15 ml dieser Lösung werden durch ein Membranfilter von 5 pm filtriert. 5,00 ml des Filtrates wird zu 500,0 ml mit 0,01 N-Salzsäure verdünnt. 10,00 ml dieser Lösung werden mit 0,01 N-Salzsäure zu 50,00 ml verdünnt.

Vergleichslösung:

Etwa 200 mg Metamizol-Natrium . 1 H_2O-Vergleichssubstanz, genau gewogen, werden zu 100,0 ml in Wasser gelöst. 5,00 ml dieser Lösung werden zu 500,0 ml mit 0,01 N-Salzsäure verdünnt .

Messung:

Die Spektren von Probe- und Vergleichslösung werden in 1 cm-Küvetten im Bereich von 320 - 220 nm gegen 0,01 N-Salzsäure aufgenommen und im Absorptionsmaximum bei etwa 257 nm ausgewertet.

Berechnung:

mg Metamizol-Natrium . 1 H_2O pro Suppositorium

$$= \frac{EP \,.\, EwV \,.\, 5 \,.\, 500 \,.\, 500 \,.\, 50}{EV \,.\, 100 \,.\, 500 \,.\, 5 \,.\, 5 \,.\, 10}$$

$$= \frac{EP \,.\, EwV \,.\, 5}{EV}$$

EP = Extinktion der Probelösung
EV = Extinktion der Vergleichslösung
EwV = Einwaage der Vergleichssubstanz in mg

Figure 4

LENGTH OF CLAUSE (AS THE QUOTIENT OF NUMBER OF WORDS PER SENTENCE DIVIDED BY THE NUMBER OF FINITE VERBS PER SENTENCE) IN 12 SENTENCES (S_1 - S_{12}) IN THE TEXTS T_1 DE, T_1 SP, T_1 EN

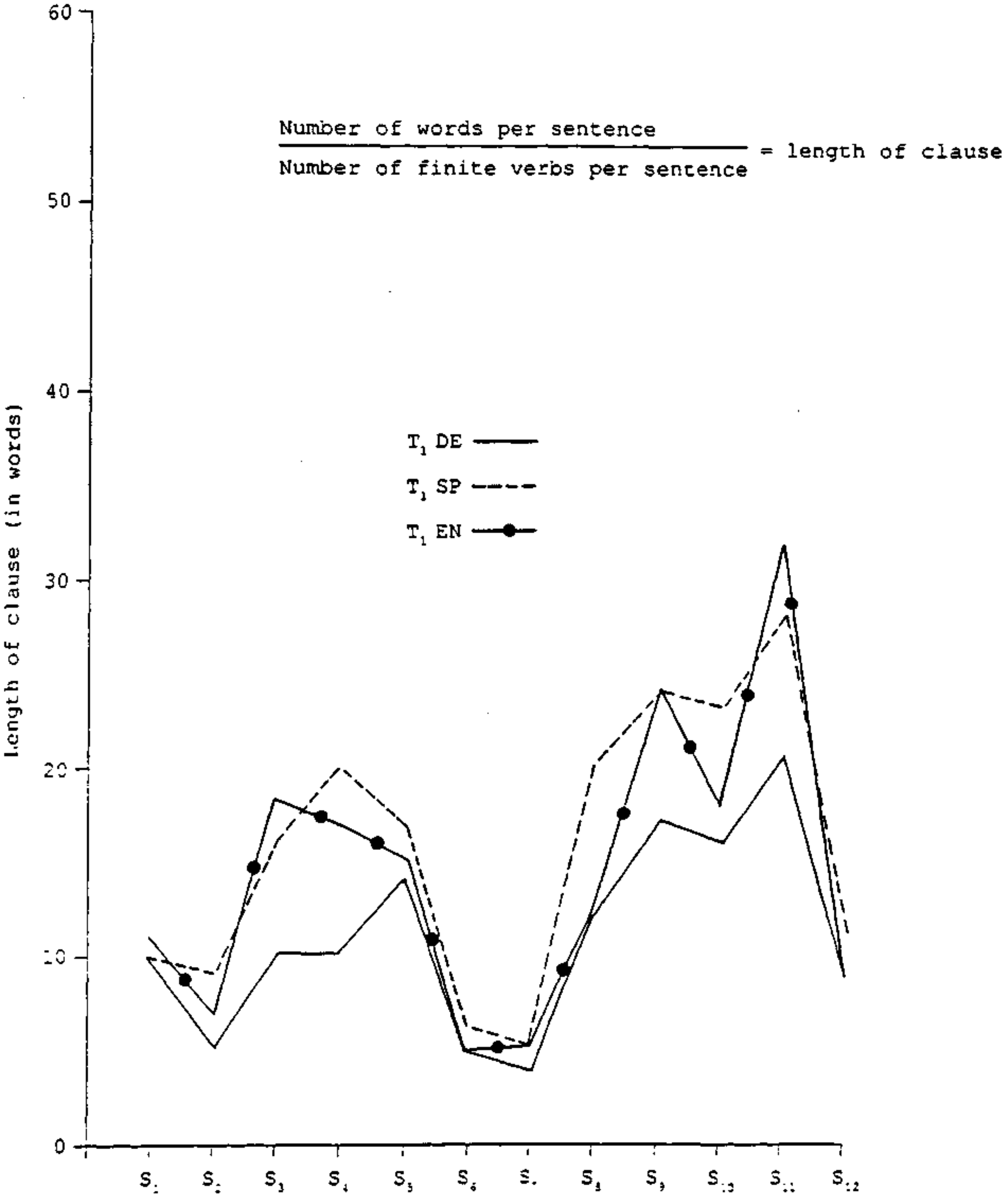


Figure 5

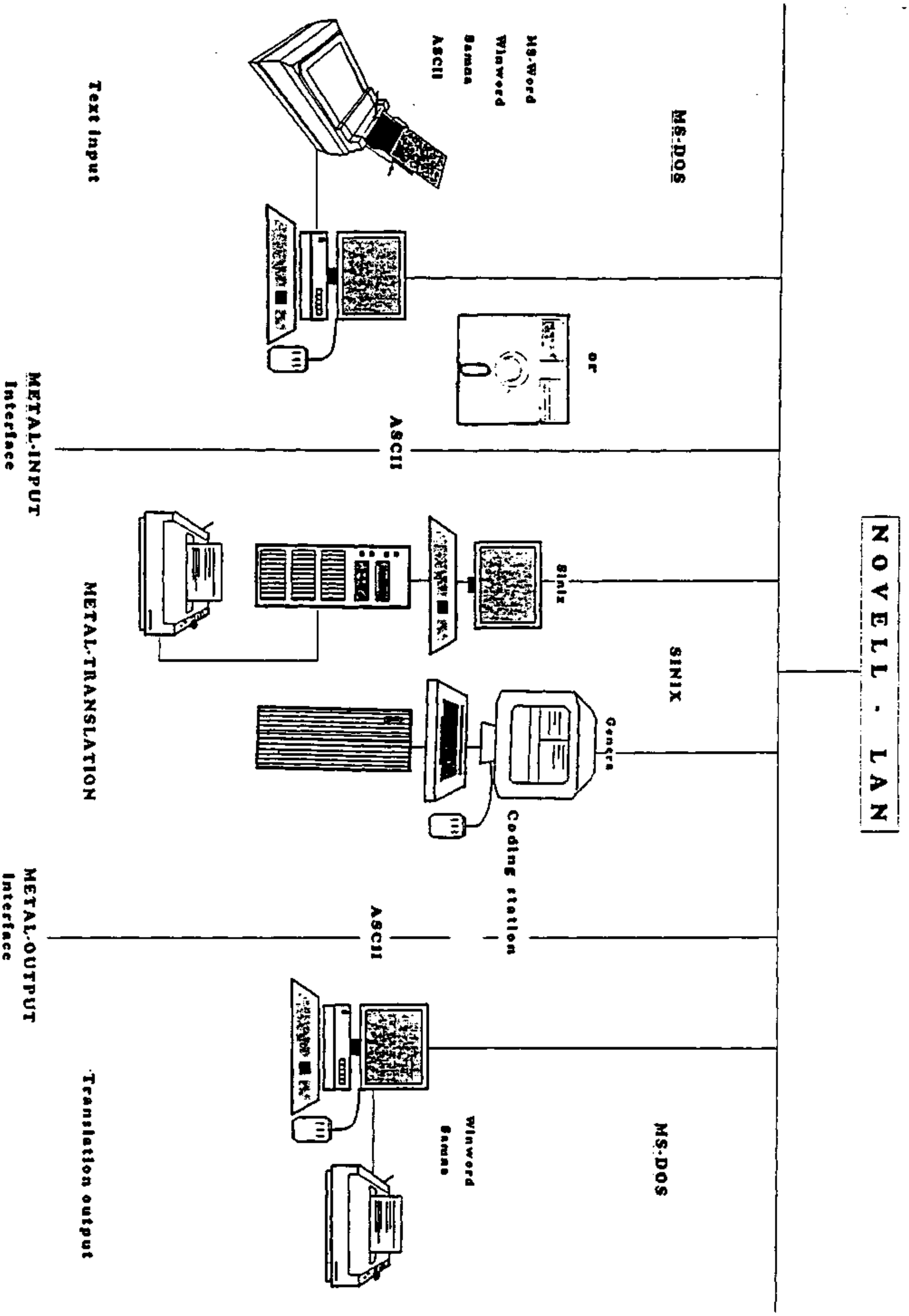
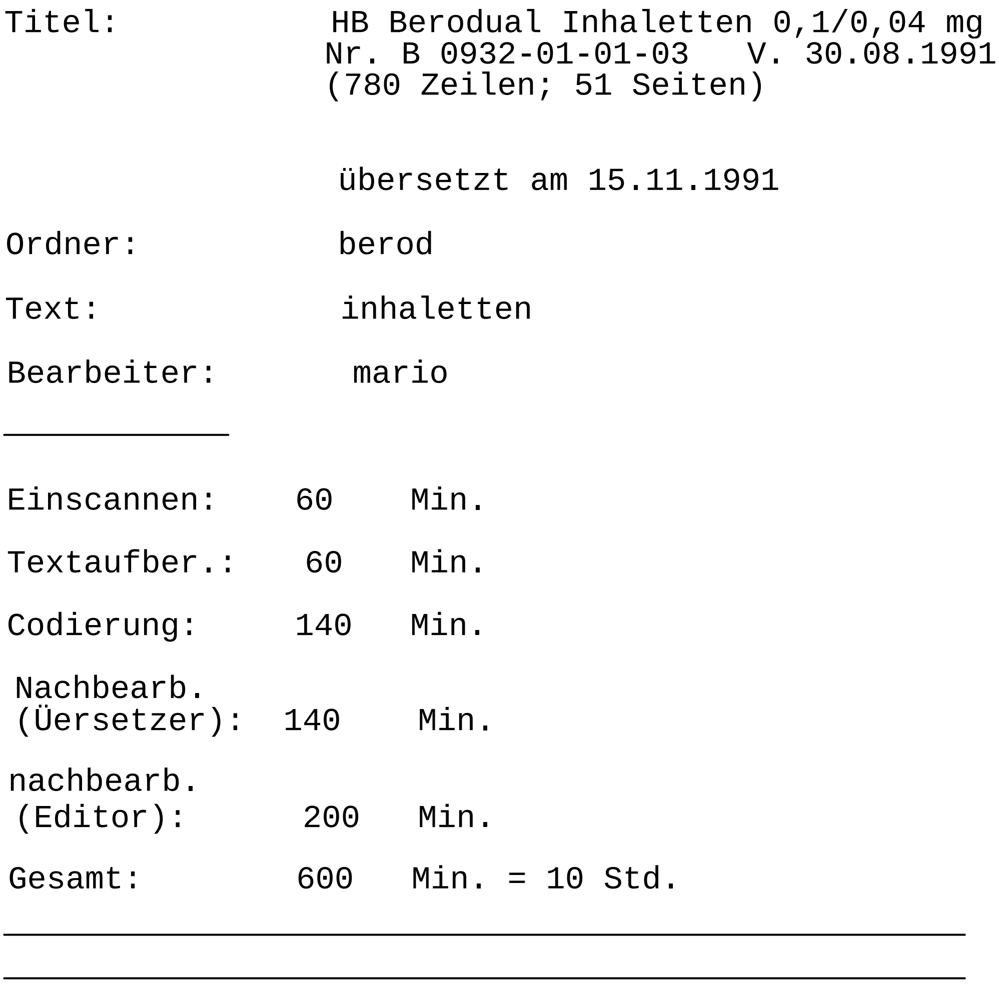

Text input
MS-Word
Winword
Samna
ASCII
MS-DOS
or
ASCII
METAL-INPUT
Interface
Sinix
SINIX
NOVELL - LAN
METAL-TRANSLATION
Genera
Coding station
ASCII
METAL-OUTPUT
Interface
Winword
Samna
MS-DOS
Translation output

Figure 6



Titel: HB Berodual Inhaletten 0,1/0,04 mg
Nr. B 0932-01-01-03 V. 30.08.1991
(780 Zeilen; 51 Seiten)

übersetzt am 15.11.1991

Ordner: berod

Text: inhaletten

Bearbeiter: mario

Einscannen:	60	Min.
Textaufber.:	60	Min.
Codierung:	140	Min.
Nachbearb. (Üersetzer):	140	Min.
nachbearb. (Editor):	200	Min.
Gesamt:	600	Min. = 10 Std.

Figure 7

101.

DEUTSCH - ENGLISCH

Titel:	Wirksamkeit und Verträglichkeit von 100 mg rt-PA i.v. in 90 Minuten bei Patienten mit akutem Myokardinfarkt (U89-0221) v. 13.10.1987 (Tabellen = 89 Seiten) übersetzt am 07.01.1992
Ordner:	retiene
Text:	rtpa2; rtpa3; rtpa4
Bearbeiter:	alexa

Einscannen:	1,5 Std.
Textaufber.:	3 Std.
Codierung:	7 Std.
Nachbearb. (Übersetzer):	4 Std.
nachbearb. (Editor):	14 Std.
Gesamt:	29,5 Std. = ca. 4 Tage

Figure 7 a

Titel:	HB Berodual Dosieraerosol 10 ml/200 Hübe à 0,020 mg SCH 1000 BR und 0,050 mg TH 1165 A (mit 0,2 % Soja-Lecithin) Nr. B 0930-11-01-04 v. 10.04.1992 (1220 Zeilen; 69 Seiten)
	übersetzt am 29.06.1992
Ordner:	berodual
Text:	berol - bero3
Bearbeiter:	mario

Textaufber.:	80 Min.
Codierung:	100 Min.
Nachbearb. (Übersetzer):	130 Min.
Postediting (Editor):	520 Min.
Gesamt:	830 Min. = 14 Std.

Figure 7b