PaTrans - a MT-system
Development and Implementation of and Experiences from a MT-system

Viggo Hansen
Lingtech A/S (Denmark)

Abstract

1. Background

Lingtech A/S (A/S = Inc.) was established by two Danish major patent attorney companies, Hofman-Bang & Boutard A/S and Lehmann & Ree A/S.

The background for the formation of the company was the introduction of the European Patents agreement the effect of which was predicted to give a considerable increase in the number of European patents to be validated in Denmark, and consequently had to be translated into Danish. To be able to meet the increased translation work-load, the two companies decided to use any available modern technology to assist in establishing the translation facilities needed. It was further decided to do so in a joint venture to be carried out through Lingtech A/S, a new company formed for the purpose.

2. Market Research

A thorough research of available translations systems on the market proved that none of the existing systems could be used for the purpose. Either they were word to word translating systems or they required the presence of an operator during the translation process.

Our need was a system which could work as a "translation factory", i.e a continously process of feeding in documents in English resulting in fully translated documents into Danish.

It was recognized that the European Commission had considerably translation problems working with ten official languages, and an approach was made to a Danish member of the European Parliament. Through that source we received documentary material stating the level of machine translation efforts carried out by the EEC. It was understood that EEC ran a project EUROTRA under the supervision of the Danish professor Mrs. Bente Maegaard, who in addition was the manager of Center for Language Technology (CLT), an institution under the Danish Minister of Research and Development.

An approach to professor Maegaard subsequently resulted in a series of meetings the conclusion of which was that a cooperation between CLT and Lingtech to further investigate the possibilities of developing a purpose-designed system was started.

3. Feasibility Study

The project was further described and it was decided that the test area should be English - Danish translation of patent documents related to petro-chemistry. A number of patent documents was selected to form the base from which the system requirements should be formulated including vocabulary, morphological-, constituent- and relational-structure.

The study should take its platform from the existing EUROTRA-project and result in a recommendation as to the feasibility of implementing a "factory-style translation system".

The study lasted about nine months. During that period the system requirements was modified according to the disclosure of actual needs.

4. Conclusion

The study concluded that it was possible to construct a system meeting the requirements as described in the background material for the feasibility study.

Consequently Lingtech decided to prepare a detailed contract proposal as background for negotiating a contract on the actual development of the system.

The contract included

- a project description according to the feasibility study
- a specification of system requirements including system performance as to speed and accuracy of translations
- a time schedule
- any other legal rights and obligations of the two parties

Furthermore the contract stated that detailed system specification of the modules forming the system should form part of the contract when such specifications - according to the time schedule - were prepared. Likewise, a schedule for intermediate tests was made and formed part of the contract.

The total development costs were established and included in the contract.

On December the 20th, 1991 the contract was signed by the two parties.

5. Development Process

Following the contract signing a Project Management Team with representatives from as well CLT as Lingtech was formed to supervise the development process.

A number of CLT working groups was initiated and the development process began.

The project comprises the following major system modules:

PaTrans - Translation module including

PaTrans lingware
(grammatical coverage, establishment of general English and Danish terms)

PaTrans software
(term- and translation rules identifier. The analysis process including the morphological analysis, the constituent structure analysis, the relational structure analysis, the interface structure analysis transferring to the synthesis phases including the same steps in reverse order)

PaTerm - Term coding module
(tool for coding of words and multiple word terms, version 1 and 2 for experienced and non experienced users respectively.)

PaTerm introduces a new and revolutionary concept of coding words and terms. The program is dynamic functioning, i.e. it is self-modifying according to the actual coding structure. When a term is to be coded, the program performs an analysis of the word and suggests inflection rules, composition rules etc. accompanied by examples of outcome of any chosen option. The program makes it possible for an non-linguistic expert to do dictionary coding. To facilitate the expert user of the program a version 2 is developed enabling the linguistic expert to use short-cut mode of operation excluding examples and other program assisted aids.

PaEd - Pre- and post-editing module
(tool for preparing documents prior to translation and editing translated files)

The modules mentioned above in general will get the system to work as described below:

1. PaTrans assumes that the text to be translated is equipped with SGML-codes to identify document name, dictionary/ies to be used, headings, style etc. Before a translation is initiated, it has to be decided which dictionaries to be used during the translation.

Dictionaries

PaTrans is operating with a **general dictionary** including the most frequently used ordinary terms of

the english and danish language, i.e.verbs like be, have, will, shall, may and noun like man, house, wall and machine.

To facilitate the proper translation of a word that can have different meanings like for instance bank, engage, composition and agent a number of domain-specific dictionaries are established. The dictionaries can be considered as organized in a tree structure but are not limited in their use to said structure. For example: A text relating to petro-chemistry is to be translated. The operator defines for PaTrans that said text should use the following dictionaries: 1. petro-chemistry, 2. general chemistry, 3. patent terms. That will cause PaTrans to look for words first in the dictionary for petro-chemistry. If PaTrans fails to find the word in that dictionary, PaTrans will look for the word in the dictionary for general chemistry. Failing to find the word there PaTrans will look in patent terms and finally in the general dictionary.

In other words, the operator initiating the translation directs PaTrans to search the words in specific dictionaries in a given sequence.

2. PaTrans is looking up all words in the dictionaries as specified by the operator and is preparing a report on its findings. In this context words also could be phrases consisting of two or more words, as for instance the English term lubricating oil' which appears as two words but translated to Danish appears as one word. Normally the report is used only for identifying the number of unknown words, but on request a full report including word-frequency, dictionaries used for individual words etc. can be produced.

   Now the operator can decide either to terminate the procedure and have missing words coded, or can decide, that the translation will continue.

3. PaTrans recognizes sentences as the object for translation. In principle each sentence (defined as everything between two periods) is an independent piece of text. Each sentence will internally be numbered and translated sequentially.

4. PaTrans performs a detailed analysis on the morphological structure of the words in the sentence. From now on each word will be present in its basic form, i.e. 'was' will be 'be′ plus a chain of information about 'be' as used in the sentence.

5. PaTrans performs a detailed analysis of the constituent structure of the sentence during which a tree structure of the words in the sentence will be established according to their interrelations.

6. Before PaTrans concludes the English part of the translation process PaTrans performs a analysis of the relational structure an in depth grammatical analysis of the sentence. The sentence is now presented with all words in their basic format sequenced according to their importance.

7. PaTrans now makes the transfer from English to Danish by simply replacing all English words with their Danish equivalents.

8. -Now Patrans will repeat the steps done so far in reverse order and according to rules valid for the Danish language, and from the analysis done in the steps 4 to 6 get to the syntheses regarding the relational structure, the constituent structure and the morphological structure.

9. PaTrans finally do the relevant inflection of the words in the sentence, and the translation is done.

10. If PaTrans meets words unknown to the system said words will be specially marked. PaTrans however will analyse the word to find out what type of word it might be (verb, nom etc.) in order to ensure a correct grammatical analysis of the sentence as a whole. Even in case where PaTrans is meeting word compositions unknown to PaTrans the system will continue the translation process. PaTrans will however associate a so called 'fail-soft'-mark with the sentence to be easily recognized by the proof-reader.

During the course of the development process detailed system descriptions were constantly updated and agreed upon of the two parties to ensure maximum accordance between expectations and outcome. The above mentioned steps 1 to 10 represents the final outcome of the functioning of the system.

Interval follow-up and -tests ensured that the process was running according to time-schedule and included such elements as formulated in the contract.

6. Hardware Selection

Since the system was a development on the EUROTRA-project it was given, that the system should work under UNIX system control.

Bidding documents including detailed description of operational conditions, program requirements etc. were prepared, and two computer companies were invited to give their bid.
Following as election process HewlettPackard was selected as hardware supplier. (HP9000/715 work stations).

7. Operation Environment

As earlier mentioned Lingtech's concept of operating is the "factory-style". You can say Lingtech is a translation factory.

Incoming documents for translation are fed into the computer via a OCR-scanner. Documents are preparsed and any word unknown to the system is reported for coding.

The document is pre-edited, i.e. via the program module PaEd peculiarities in the document are identified. It could be non-translatable words such as trademarks, or marking of style and size of letters, or identification of headings or tables wherein a word-to-word translation is requested. Following the pre-

editing and when word/term-coding necessary for a successful translation is done, the translation of the document is initiated.

The system makes the translation and stacks the translated document. Via the control panel of the main computer operational control can be maintained over all work-stations. The translation-program does never give up. If the system recognizes that a given sentence is so constructed that it is unknown to the rules accepted by the program, a detailed analysis of each word in the program results in a suggested translation. The sentence in question will be marked (fail-soft). If the system recognizes that more than one possible translation can be relevant alternatives are hidden under the selected translation and can be called upon on request.

All machine-translated documents are proof-read by terminology experts. Via PaEd a post-editing function is initialized. The proof-reader calls the document on the computer videodisplay. The document will be read sentence by sentence. Machine marked fail-softs are given special consideration and any correction is done. If deemed necessary the proof-reader by mouse-click on the actual translated sentence in a separate window can have the untranslated equivalent PaEd assist the proof-reader in bringing back the document in its original lay-out The process can then be terminated by either outprinting or diskfiling the translated document

8. Operating Costs and Cost Savings

The system has proven to save substantial costs in the translation process.

The table below shows the relevant cost figures related to the translation process, and it will be apparent from the table that actual translation costs are more than halved by using the MT-system.

Cost Performance per 2 million translated words per year

	MT-environment		***Manual translation***	
	DKK	***US$***	***DKK***	***US$***
HARDWARE				
Scanning/Pre-editing	6000	1000	6000	1000
Workstation	30000			5000
SALARIES				
Scanning/Pre-editing	84000	14000	12000	2000
Manual Translation			132000	220000
Proof-reading	420000	70000	330000	55000
Coding/System control	102000			17000
Total	642000	107000	1668000	278000
Yearly cost savings	1026000	171000		

9. Experiences After ½ Year of Operation

* Actual performance compared to requested performance

The average speed of the translation system is about twice the requested speed, mainly due to faster hardware but also due to intelligent programming.

* Proof-Reading

The necessary time for proof-reading seems to be as forecasted i.e. proof-reading takes about one quarter of the time of a manual translation.

* Actual Development Costs

During the development process a strict budget control has been exercized resulting in actual costs equivalent to the budget. Even alterations to the system during the development process could be covered within the original budget.

* Need of a TAS (Total Application Solution)

Much care has to be given to looking at the machine translation system as a tool within a whole document handling and translation process. If the document handling before and after the machine translation occupies so many resources that common sense tells that the whole process can be done faster by manual translation the concept of machine translation fails.

Therefore, it is imperative that tools are developed along with the machine translation systems to ensure that a.o. the following elements in the total process of converting a document in one language to a similar document in another language are satisfactorily covered:

- Documents should be easily read into the computer via a scanner system in such case where an electronic copy of the document is not available

- An editing system for pre- and post editing of documents has to be very user-friendly and operatable by ordinary office staff

- Drawings, formulas, illustrations etc. must be recognized and stored in the computer and automatically collated with the text in a post-editing procedure.

- Dictionary coding has to be an easy process, but also a very precise process with regard to coding of such elements necessary for correct translation both what relates to gramma, as to sentence composition and selection of right word in multiple choice cases.

- Definition of dictionaries and their priorities in the translation process is important and difficult. In

- many cases the same word can have different meanings. Originally we estimated to operate with about ten different dictionaries each representing one area of terminology. However, operation experiences have proved that a much smaller number of dictionaries makes better translations!

*** Intensive Management Control**

During the development and the implementation process an intensive management control of the whole operation is extremely important.

Common sense has to be mixed with a detailed understanding of actual cost involvement in adding facilities to the system. The process - and the money - can easily run out of your hands, if you do not set common sense limits for the degree to which your system should work 100% perfect.

The development itself must be followed carefully. Scheduled results must be obtained in the desired quality and at such time as laid out in the time schedule.

But most important of everything: As a project manager it is your responsibility, that you can explain yourself in such a way, that it precisely defines your needs, and you at the same time takes the efforts to control that your expressed needs are clearly understood by the computer- and linguistic experts in your way and your meaning. Any misunderstanding falls back upon yourself. You are to blame!

10. How to Get More Companies to Use MT-Systems

* - Standardisation of domain-specific dictionaries

* - Easy to use systems

* - Easy to understand systems

* - Ability to fit into existing procedures

* - The TAS-concept