VII. PILOT PROJECT DEFINITION.

1. Overall contents and aims of the pilot project.

The DLT pilot project is the next major step in the development of the DLT system.
It follows the current DLT feasibility study. Whereas the latter has been an exercise entirely limited to "paper" (or, expressing it more positively, "brainware"), the pilot project has to provide the evidence of a software-implemented, partial but operational DLT system.

As well known in the development methodology of general EDP systems, a certain amount of interaction and iteration is required between the successive phases of feasibility study, global design, detailed design etc. The development of DLT, a more or less pioneering and research-like undertaking, will definitely need such interaction and iteration between design and operation [we deliberately do not use the word "experiment", because of its connotations with pure research].

At the time of finishing the current feasibility study, we have the feeling that most of the relevant "paper-work" for DLT has been done (thinking started already in 1980), and that a continued effort should be accompanied by working with a computer model.

The present feasibility study serves as the justification for the pilot project, which in turn will serve as the justification for the phase to follow (a trial production system), in accordance with usual policies of step-wise investment.

The size of the pilot project (roughly 12 man-years) will be a multiple of that of the EC-supported feasibility study (about 2 man-years), even if one adds BSO's preliminary investment (another 2 man-years) to the latter. The pilot project's duration is approximately 2 years.

In order to be flexible enough to meet the variety of financial possibilities and constraints likely to come up in the wake of this feasibility study, several options in the choice of a pilot project will be presented. Therefore, we have split the pilot project into work-packages with separate price tags. These packages will be described below. Summarizing remarks and overviews will follow in a concluding section [VII.8].

2. Work Package Ia: the IL kernel.

This work package can be seen as the first and most logical continuation of the present work. It concerns the middle phase of the translation process, common to all future SL's and TL's, and centers around DLT's major characteristic: the Esperanto-based compact IL.

The IL kernel as contained in work package Ia includes Step 4 and Step 5 of the DLT process sequence [see section III.4.1], i.e. the IL-recognizer and the IL-parser (an extension of the kernel is described under work package Ib).

The contents of package Ia can be outlined as follows:

a. Completion of the IL grammar description.
The grammar of the IL, as far as described in detail in this report [Chapter IV], covers an estimated 65% of the total amount of modifications in syntax and morphology (function words, affixes, particles) from existing Esperanto [layers I and II described in section IV.1.2]. A further 25% of the grammar definition needs to be added before carrying out further sub-packages (the remaining 10% represent omissions and inconsistencies that can only be handled during and after test runs on the pilot system).
Estimated effort: 6 man-months, preferably immediately after the feasibility study.

b. Completion and programming of the ATNs defining the IL.

Whereas sub-package "a" concerns a descriptive definition of the IL, this sub-package deals with the formal counterpart. The formal IL definition in the form of ATNs will be the basis for computer implementation and must therefore be considered the most genuine record of the IL grammar (the descriptive form should in fact be considered as a derived form of definition, even when it is completed before the formal one).
Departing from the present form of the IL-ATNs [section IV.3.2 and appendix pp. 8-15], their completion exists in:

i. The addition of several grammatical phenomena and details: clausal coordination and ellipsis, the categories of intensifiers and modifiers, punctuation (commas), etc. Also: desk-checking of the ATNs.
Estimated effort: 3 man-months.

ii. Formal detailing and consolidation of the ATNs by exactly stating:

- conditions along arcs (including nesting depth restrictions);

- actions along arcs (register setting and IL-tree building).

In accordance with the strategy of a separate IL-recognizer and IL-parser [as explained in section of III.4.1 of this report], it should be reminded that two sets of ATNs are required. These sets largely coincide, but the one (the recognizing ATN) is void of IL-tree building actions (a syntax-checker without code-generation), the other (the parsing ATN) is void of correctness-checks (such as checks on nesting depth or chain length restrictions).
Further, it should be taken into account that the set of IL-ATNs will be used as a departure point and common backbone for the SL-ATNs to be developed in a later project phase [see also III.4.2.3.5a], which increases the importance of modularity and ease of updating for the ATN-implementations.

The dominating work sequence will be:

- adding notes along the arcs of the ATN-graphs, heavily relying on abbreviations;
- writing out the IL-ATNs in a computer file, using the high-level programming language PROLOG; the clauses of this program will contain all the ATN conditions and actions (the so-called "augmentations") in full detail; a "paperless" program development environment will be used (incl. automatic checking of PL syntax).

Estimated effort: 3 man-months.

Estimated effort ("b"): 6 man-months.

c. Finalization of the IL tree structure.
This includes:

- a comprehensive description of the handling of syntactic complexities (conjunction at all levels, ellipsis, different type of modifier chains, etc.), both with respect to their symbolic (graphical) as well as to their computer-internal representation (pointer structure);
- a complete description of the contents and format of node labels [see section IV.3.3].

Estimated effort: 3 man-months.

d. Composition of a test-set of IL sentences.

The DLT system design and development strategy aims at combining the needs of a step-wise development (starting with an operating kernel at an early stage) with the reliability requirements of a more fully developed translation system (including one or more SL-modules) at a later stage. Therefore, the IL-recognizer [see also III.4.1] will serve future integrity (according to the dual programming principle), but it will protect the DLT-network and its receivers (i.e. the IL-parser) also and especially during the time when "manually" generated IL-input still has to replace a working SL-module.

The IL-recognizer as well as the IL-parser have to be tested themselves by manually generated IL-input.
One way to do this is by submitting IL sentences in character-string or surface form, including such separators as extra spaces ("pop-spaces" etc.). A second way to do it will be mentioned in section 3 (Work Package Ib).

The test-set should be composed in such a way that all syntactic patterns and a large number of combinations of them will be covered (this implies that all arcs of the ATNs, and a large number of possible paths through the ATNs, will be tested). Also, a part of the test-set must be derived from "manually simulated" future SL-modules, in much the same way as the manual DLT translation reported about in section V.4.2 of this document. Further, a contribution from a critical person at some distance from the team is desirable.
The number of sentences in the test-set (counting simple as well as complex sentences) will be in the order of 1000. The set should include incorrect as well as correct IL-sentences, in order to test the IL-recognizer's "filtering" capability.

Estimated Effort: 2 man-months.

e. Specification of required IL-parser output (parser-generated trees) for each sentence of the test-set mentioned under "d".

A test organization with a maximum throughput and a minimum of clerical work is aimed at. The tree structure which the IL-parser SHOULD produce (the so-called "SOLL"-tree) for a given input sentence will be automatically compared with the structure it ACTUALLY produces (the so-called "IST"-tree). In this way, the whole batch

of 1000 test sentences can be processed over and over again (i.e. after each ATN-modification) with the computer sorting out the ones that are parsed wrongly. Of course, no output trees need to be compared for (deliberately) incorrect input sentences; for these, only the rejection by the IL-recognizer needs to be checked.

Estimated effort: 3 man-months.

f. Software utilities for the running of ATNs and the handling of tree structures.
This comprises system-software provisions for using PROLOG; display (tabular), storage, updating and execution of ATNs; specification, storage, manipulation, comparing and display (non-graphical) of trees, etc. Standard and portable software will be used as much as possible, but some effort for local support and assistance must be calculated.
Estimated effort: 4 man-months.

g. Construction of the IL monolingual dictionary.
This is in fact the IL column of later bilingual IL-TL dictionaries. In addition, it will contain provisions for accepting and checking manual character-string input (see under "d") and converting character strings to morphems. The work consists of:

- conceptualization and design (format of dictionary entries, range of grammatical features etc.);
- design and implementation of interactive utilities for first input, input-error screening, consistency-cross-checking (making use of redundancy deliberately designed into the dictionary entry format), updating, tabulation etc. of dictionary entries;
- filling in the dictionary (first input) via the interactive utility.

Estimated effort: 12 man-months (if no standard dictionary-utility such as COMSKEE can be used, another 12 man-months must be added for software-tool development).

h. Attaching function word compatibility matrices to the IL-recognizer.
A number of matrices have been composed [in the course of the feasibility study], which indicate the abilities of function words to modify each other, covering intensifiers, determiners, numerals etc. Attachment of these matrices to the IL-recognizer ATN (as additional conditions on certain arcs) will enhance the recognizer's

filtering capacity for incorrect (i.e. ungrammatical, illegitimate IL) input.
On their turn, the matrices themselves have to be checked on completeness and integrity, which can only be achieved in the test environment of a pilot system.
This task includes linguistic maintenance of the matrices and software provisions both for operating and (easy) maintenance.
Estimated effort: 4 man-months.

i. Testing proper of the IL-recognizer and IL-parser (the "IL kernel").
Locating errors in the ATNs. Filling gaps and making improvements in the grammar and the ATNs. Evaluation, documentation and reporting.
Estimated effort: 5 man-months.

Estimated total effort (Ia): 45 man-months.
Estimated critical path length: 17 months.

3. Work Package Ib: the extended IL kernel.

Instead of departing from (manually prepared) IL-string input, the pilot system kernel can be extended to depart from IL-trees. This means that the IL-kernel will be extended towards the SL-analysis side, where it will then include Step 2 and Step 3 of the DLT process sequence [see III.4.1], covering: Tree Ordering, Separator Insertion, Agreements Regulation and Tree-to-String Conversion.

The IL-trees forming the input interface of the extended IL kernel are the trees produced as output of Step 1 of the DLT process sequence: the IL-directed SL-analysis. Because SL-analysis (which is the most expensive DLT component) will not be realized within the pilot system discussed here, the IL-trees have to be prepared and entered manually.

Work package Ib consists of:

j. Programming and testing of the algorithms for Tree Ordering (according to the canonical IL word order), Separator Insertion (both "pop-spaces" and "skip-spaces", ATROP-mechanism [see section IV.3.4]), Agreements Regulation (agreement of case and number between noun, adjective and determiner endings) and Tree-to-String Conversion. The latter will produce an unbracketed linear string of IL (Esperanto) morphemes, each of which can be made externally visible as a character string for easy inspection. The morphem string is passed to Step 4 and Step 5 [work package Ia above] in an internal code (the

compactness of which is crucial in the eventual but not in the pilot DLT system).
Estimated effort: 8 man-months.

k. Composition of a test-set of IL-trees and evaluation of the tests.
This task includes the specification of the corresponding IL-strings that are to be produced as output after Step 2 and Step 3 [cfr. task "e" above], in order to facilitate a quick test procedure. It also requires insight into the prospective products of Step 1.

Estimated effort: 5 man-months.

Estimated total effort (Ib): 13 man-months.

4. Work Package IIa: TL-module for German.

Though the IL kernel (treated under Work Package I) is the heart of DLT and the common pivot in all its future translation processes (largely determining the overall performance, maintainability and extendibility of the system), DLT's practical value can only be demonstrated and assessed by carrying out language translation.

It has been pointed out [see sections III.4.2 and III.4.3] that DLT's SL-modules are more expensive to develop and require a larger research investment than its TL-modules. Also, DLT was originally proposed [Witkam, 1981a] as a "half" MT system, a system for IL-to-TL translation only, in which different TL-modules connect information consumers to a central "multilingual" (i.e. IL-) database (irrespective of how the latter would be created). Apart from a transitional development stage towards a full DLT system, such a "half" configuration remains a valuable proposition for certain applications.
Further, it can be said that the design of TL-modules in DLT is less dominated by DLT-specific characteristics than that of SL-modules (with their IL-directed intervalwise parsing and interactive disambiguation). This opens the door for cooperation with MT groups elsewhere, to make use of existing components (such as dictionaries) for the construction of a TL-module.
A phased realization in which an early phase consists of the IL kernel and one or more TL modules (the so-called "right half" of DLT) is therefore a practical proposition.

As for the number and choice of TL-modules in the pilot system, the following considerations take part.
Just one TL-module may not be a convincing proof of a

"multilingual data base facility". On the other hand, only a limited learning-curve effect on the development of additional TL-modules can be expected, because the major effort is in the TL-specific domain (with dictionary building as the critical path).
Of course, there is the commercial and practical relevance of a language [see also V.3.2], not unaffected by the country or interests of prospective investors. Another point of practical importance is the possibility to cooperate with other MT-groups and to use existing components (dictionaries, special software tools) for the TL-module of the chosen language.
Regarding dictionary size, a range of 5000 to 10000 entries of basic (i.e. general-purpose) vocabulary will be aimed at.
Furthermore, a highly flexional language (as German) will give more information about the translation quality than an inherently ungrammatical and ambiguous language like English.

Taking all this into account, German is proposed as the first TL-module (with Dutch as a second choice, in work package IIb). The following estimated effort is based on the assumption of a take-over of existing components of the MT-system SUSY of Saarbrücken: the Esperanto-German Transfer Dictionary (4000 entries), the German Synthesis Dictionary (11000 entries) and the dictionary-building tool COMSKEE. These dictionaries have to be adapted and extended to fit into DLT [without these dictionaries, roughly 30 man-months would have to be added; the tool COMSKEE represents another 12 man-months development effort].
The work (package IIa) consists of:

a. Design and coordination of the TL-module, especially with regard to the needs and practical methods for dictionary adaptation and extension.
   Estimated effort: 7 man-months.

b. Software assistance, including installation, conversion and maintenance of existing components as well as the provision of additional tools and facilities.
   Estimated effort: 8 man-months.

c. Writing the grammar for the IL-German transfer and synthesis.
   Estimated effort: 6 man-months.

d. Adaptation and extension of dictionary entries.
   This involves orientation onto the IL (instead of common Esperanto), in particular the rewriting and addition of valency information in connection with IL-typical function word usage.
   Estimated effort: 9 man-months.

Estimated total effort (IIa): 30 man-months.
Critical path length: 16 months.

5. Work Package IIb: TL-module for Dutch.

In the Netherlands, no MT-systems have been built in the past (not even experimentally), nor have Dutch TL-modules of any substantial size been tried out in MT-systems elsewhere. This means that practically no usage of existing components can be reckoned with.
On the basis of earlier estimates [in 1981], a comparison with package IIa, and the learning-curve effect, we get:

Estimated effort of IIb (if following IIa): 40 man-months.
The work tasks will correspond to those described under IIa (with 'adaptation', 'extension' etc. to be replaced by 'creation'), with the addition of a 6-month task for building the Dutch Synthesis Dictionary.
Critical path length: 18 months.

If IIb is built as the first TL-module, the estimated effort is 52 man-months and the critical path length 21 months.

6. Work Package III: International Business and Law Terminology.

This package can be included in the pilot system for the following reasons [see also section V.3.2 of this report]:

1. It will enrich the pilot system with an additional, up-to-date and business-application oriented vocabulary. This will bring the pilot system nearer to the eventual production system.

2. Postponing this work till after the pilot system could unfavorably affect the overall critical path length towards a DLT production system.

As to the contents of the work, two major components can be distinguished:

a. Working out a basic version of a Glossary of International Law Terms, with definitions in Esperanto, and translations of the terms in English and German.
This work is not DLT-specific and can serve other purposes (publication) as well. It consists of the checking, collecting and redefining of terms, using such standard

works and sources as the PIV, the International Business Dictionary in 9 Languages [Munniksma, 1975], the Internacia Jura Revuo, etc.

Estimated effort: 12 man-months. Despite the fact that the character of the work is not DLT-specific, the effort is counted here as a Work Package for the funding and progress of which BSO will be responsible.
Estimated critical path length: 18 months.

b. Entering the Glossary into the DLT pilot system, as additional dictionary entries of the German and (future) English TL-modules. At the same time, the Esperanto definitions are rewritten into IL and entered as additional test material.

Estimated effort: 12 man-months.
Estimated critical path length: 12 months.

Estimated total effort (III): 24 man-months.
Critical path length: 30 months.

7. Work Package IV: more detailed study of SL-module, with dialogue simulation.

The SL-analysis is known to be the hardest nut to crack in an MT-process. This is also true for DLT, which features a new and ambitious combination of fast automatic parsing and human clarification of input sentences [see section III.4.2 for an impression of the SL-analysis design].

Construction of a working SL-module is considered beyond the scope of the pilot system. However, the construction of an SL-module should not become an overwhelming critical path after the pilot project. In order to have a balanced platform for further development at the end of the pilot project, a detailed elaboration of the particular parsing process chosen and a practical investigation of the disambiguation dialogue (on a working simulation model) seem justified.

Estimated effort: 24 man-months.
Estimated critical path length: 12 months.

8. Final Recommendation.

Figs. VII-1 and VII-2 give a schematic overview of the Work Packages (WPs) and the Pilot Project's overall schedule.

WP I, the IL-kernel (with or without extension Ib), is the cornerstone of the project and cannot be omitted.

WP IIa, the German TL-module (with the use of existing components from Saarbrücken), stands out as the most attractive combination with WP I.

The various options are caused by trading off team size against project duration. For the combination of WP I and WP IIa, a 4 or 5 man team can result in a 33 months, a 6 man team in a 22 months pilot project duration (not accounting for contingencies, vacations, etc.). The total effort in this combination is 75 (without Ib) or 88 (with Ib) man-months.

As to team staffing for the various WPs, computational linguists, lexicologists and other specialists (Esperantologists) will take part in addition to software professionals. In the estimated-effort figures, also the anticipated contributions of outside advisers have been included.

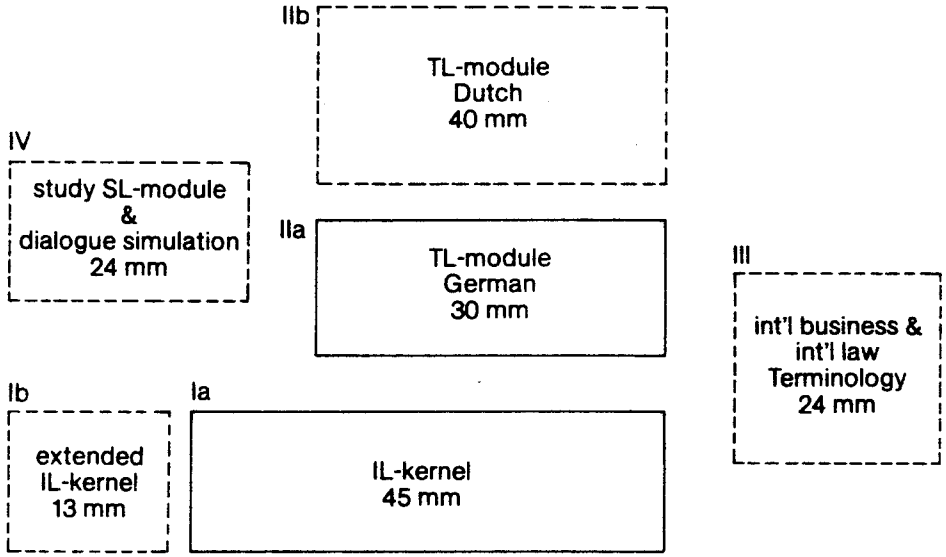
As an additional work package [not included in figs. VII-1 through VII-3], which must evidently be included in any variation of the pilot project, one can regard the installation and testing of the microprocessor prototype hardware proposed in section VI.3 [and shown in fig. VI-6a]. As has been indicated there, this will involve an extra 6 man-months.
Initial development of all software will of course take place at a larger computer facility (viz. VAX with UNIX), such as described in section VI.4.2.

Taking into account both hardware and manpower costs, the approximate expenditure for the pilot project is estimated at (in Dutch guilders, excluding VAT, price level of 1983):

- Dfl. 2.4 million for WPs Ia and IIa;
- Dfl. 4.0 million for all WPs except IIb;
- Dfl. 5.1 million for all WPs.

Finally, fig. VII-3 gives an overview of total DLT development cycle, of which the pilot project is only a part. The next stage will be the development of an SL-module, for which French appears to be the most likely candidate now (on several grounds: it is one of the primary EC languages; for MT, it is not as highly ambiguous as English).

Fig. VII-1. Pilot project WP constellation:

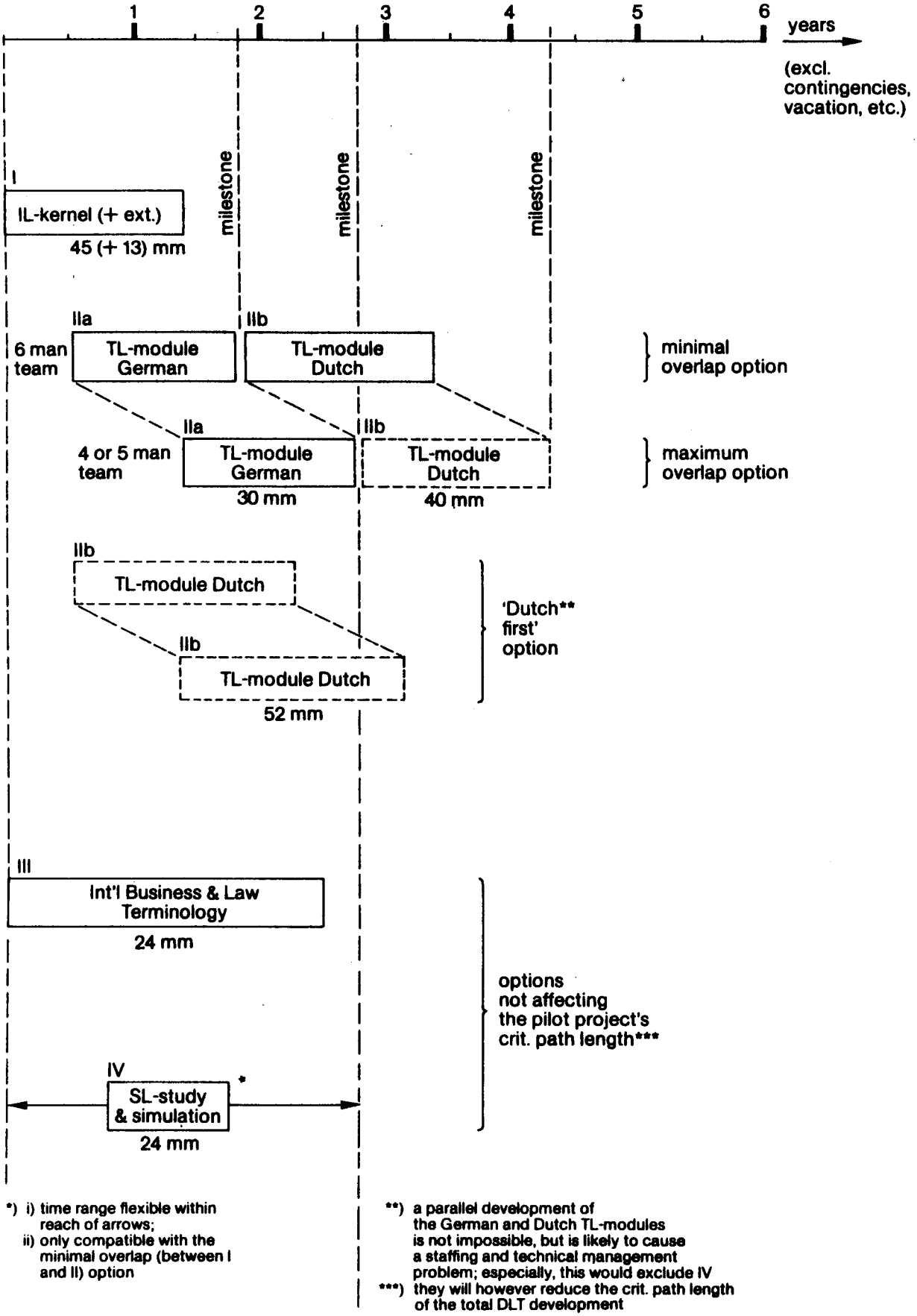


The figure indicates the estimated effort in man-months (mm), to which the sizes of the boxes are proportionate.

The solid boxes represent the Work Packages (WP's), primarily recommended for the pilot project. The dashed boxes indicate optional additions, to fit various financial or priority schemes.

Fig. VII-2.

Pilot project overall schedule:



*) i) time range flexible within reach of arrows;
ii) only compatible with the minimal overlap (between I and II) option

**) a parallel development of the German and Dutch TL-modules is not impossible, but is likely to cause a staffing and technical management problem; especially, this would exclude IV

***) they will however reduce the crit. path length of the total DLT development

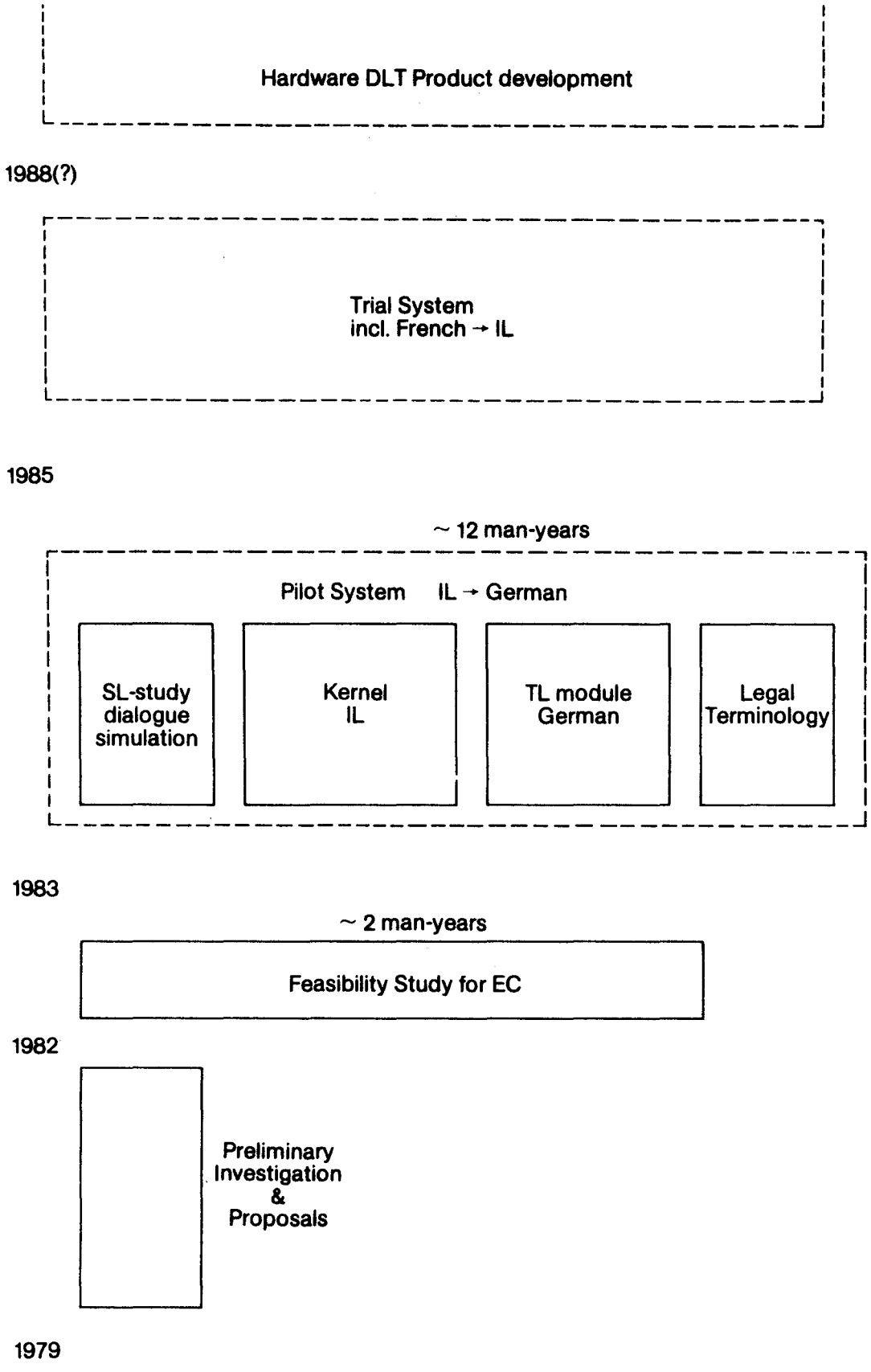


Fig. VII-3. Overview and approximate schedule of DLT development.