II. DLT MAIN CHARACTERISTICS (THE 'OUTSIDE' OF DLT).

1. First characterization.

DLT is a system for translation between natural languages (English, French, German, Italian, ...), to be embedded in computer networks and terminals. It consists of:

a. special equipment and human interaction at the sending terminal;

b. special equipment at the receiving terminal;

c. a special interface standard between these terminals.

The system permits text to be entered in (for example) English at one terminal, and subsequently to be displayed in French at another (possibly remote). A third terminal might present the same text in German, a fourth one in Italian, etc.

The translation process is in fact distributed over the network: one part takes place at the sending terminal, where the person who enters the source text also has to add some text clarifications, in a computer-initiated dialogue [see fig. II-1a]. The other part of the translation takes place upon reception in the receiving terminal, completely automatically and unnoticed: only the translated text appears at the display screen there [fig. II-1b].

Text entry (including editing), transmission and display will be handled by the usual word processing and data communications facilities. The language translation must be regarded as an optional extra service, compatible with general terminal and communication interfaces.

2. Product environment.

Originally, DLT has been conceived for <u>international videotex information retrieval and information distribution</u> systems [Witkam, 1981a]. Especially in Europe, but also in other regions of the world, a future rise of public videotex information systems together with satellite TV may create new language barriers that have to be resolved. This includes subtitling of news reports, interviews, documentary films etc.

If one thinks of the various forecasts made in the late 1970's, the market penetration of public videotex systems has clearly fallen below expectations. Though the prediction of

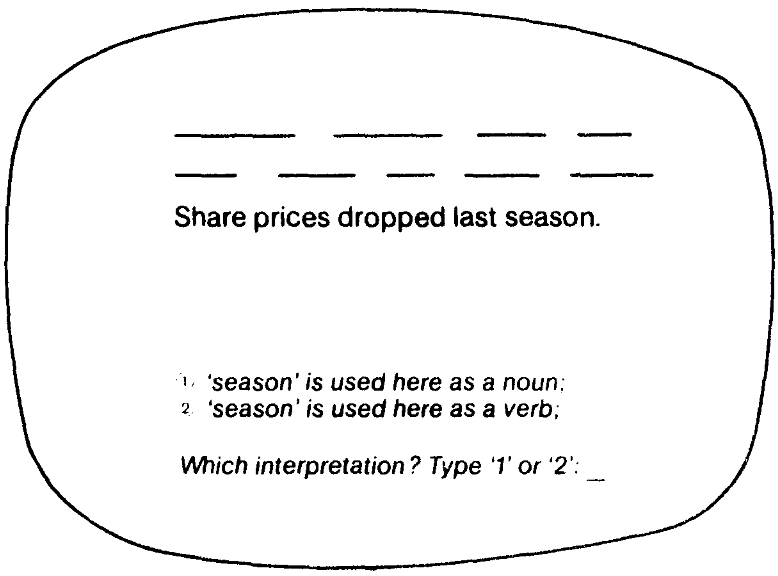
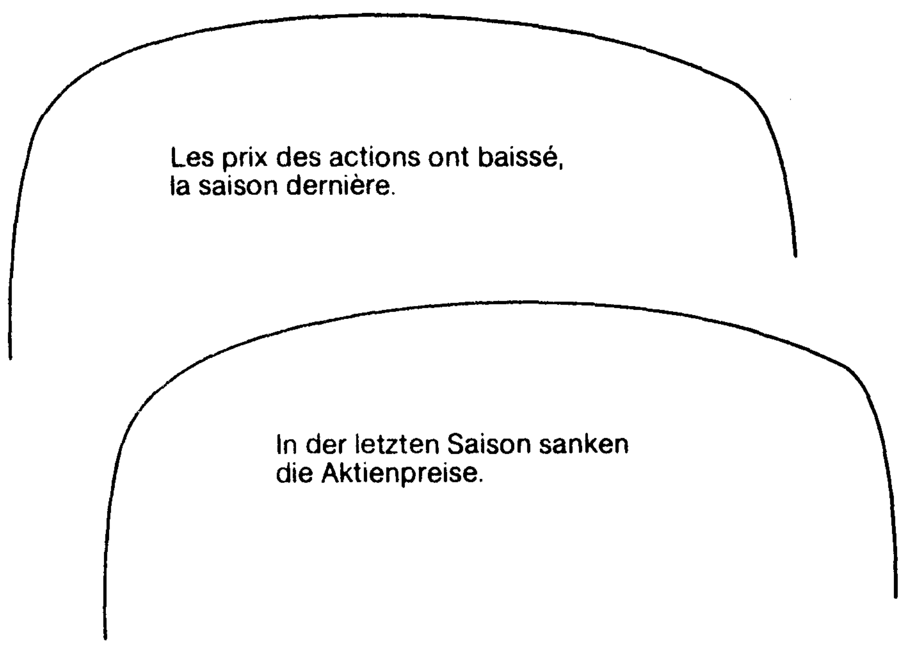


Fig. II-1a. Text entry under DLT. After a sentence has been typed, the system (italics) interrupts the typist with a request for clarification.

Fig. II-1b. Automatic display of translated text at receiver terminals.



videotex penetration appears to be most complex, one present scenario (referred to as 'Minimum Scenario' in [Maurer, 1983]) still indicates 4 million two-way videotex sets installed in the EEC in 1990. Anyhow, public videotex represents a potential mass consumer market that may rise suddenly and expand even more quickly than the development of DLT itself.

In the future videotex mass consumer market, but also in the more near and partly already existing domain of professional on-line information retrieval, the emphasis is on the receiving of information. Though the user interacts by sending an information request, the main stream of data (abstracts or full text) is towards him.
On the other side, the IP (Information Provider) generates text for a multitude of customers.
This situation permits relatively low-cost text receiving equipment at one side, as opposed to relatively high-cost text generating equipment at the other side. The DLT design capitalizes on this balance.

Two other key-words characterize the environment in which DLT will operate: OOF (Office-of-the-Future) and PC (personal computing.
In the OOF, desktop terminals will more and more replace paper trays. Electronic storage and transmission of information over LAN's (Local Area Networks) will be commonplace. For an international or multilingual staff, the provision of such a network with DLT is an ideal addition: within the supported set of languages, anybody can enter as well as read documents in his own language [see fig. II-2].

The entering of text will take place on WP (Word Processor) type of equipment. Text entry on WP's has become a normal practice in today's office. In an increasingly automated world, it is a process in which human activity is required, and this will probably remain so for a few decades. Even when speech input will catch on, human guidance and correction will be an indispensable part in the total text-entering process. DLT takes advantage of the presence of a WP operator, to restrict the cost of human assistance in the translation process. This process, or more exactly the part of it within the text-generating terminal is semi-automatic. The idea now is to use the same person both for usual WP tasks (typing, editing etc.) and for the addition of text clarifications at the computer's request (the so-called 'disambiguation dialogue' [see fig. I-1a]).

Text entering under DLT does NOT require the presence of a translator at the WP, and DLT is certainly not a tool for human translators. The latter is covered by so-called CAT (Computer Aided Translation) systems, of which the Weidner

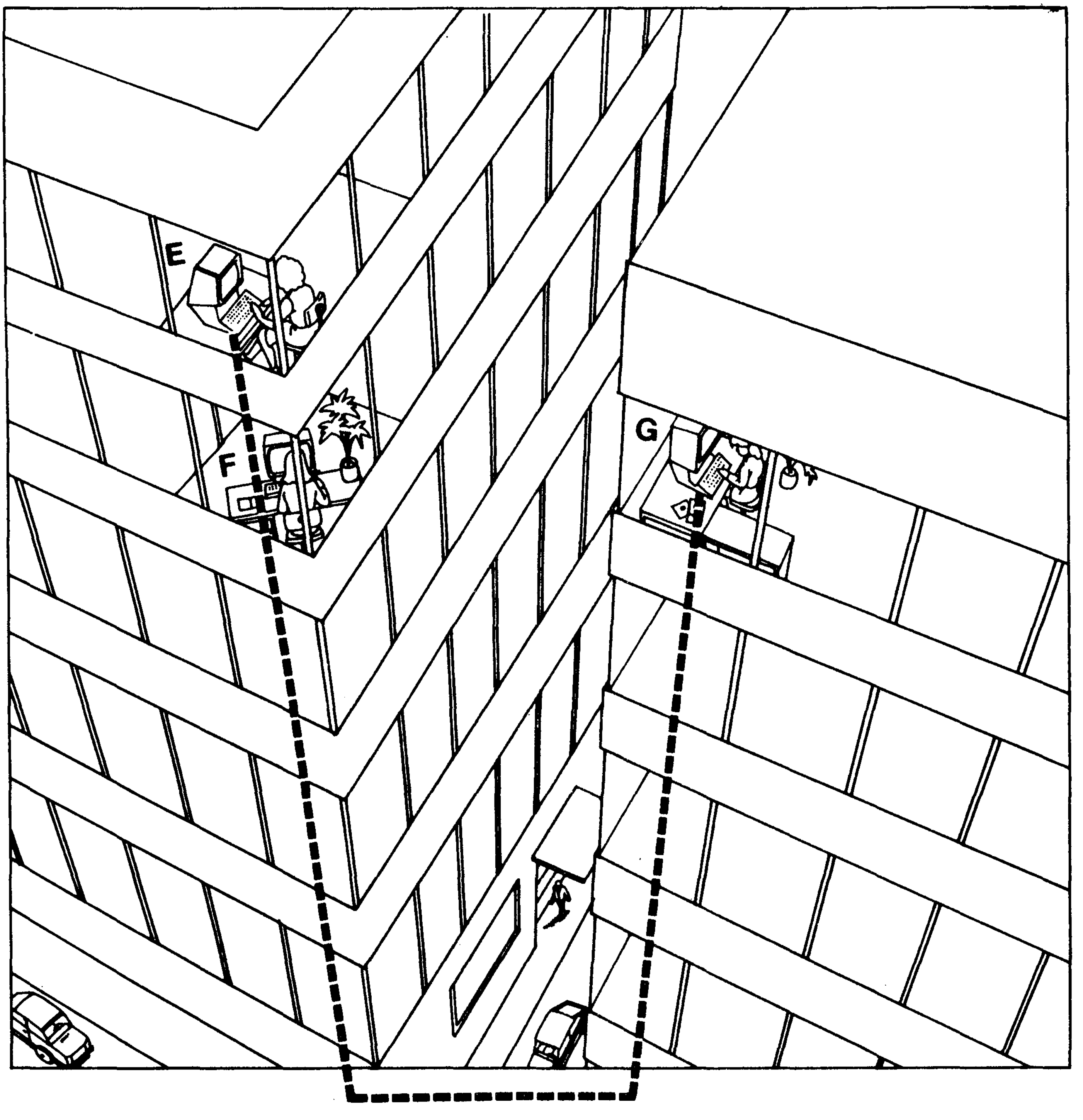


Fig. II-2. In the multilingual office-of-the-future, with a desktop terminal in every room, DLT will enable everybody to access textual information in his own language. Here for instance, a document entered in English in room E, will be transmitted over the local DLT network and appear in French at the terminal screen in room F, or in German in room G.

system has become the best known in recent years [Hundt, 1982].
At the text generating terminal, DLT only requires knowledge of the source language and understanding of the context or subject. When for instance the word 'bank' appears, DLT may ask the human operator for help, and he or she should be able to decide which sense ('side of a river', 'financial institution') applies. Sometimes, also basic grammatical concepts ('verb', 'noun') will occur in the man-computer dialogue [fig. I-1a]. By and large, the 'disambiguation' work will be within reach of the well-educated secretary, who may experience it as a task enrichment over conventional typing work. For highly technical texts, the author himself will be the most appropriate person.

Text entry by the author is of course a familiar phenomenon in PC applications, either in the office or at home. If done for hobby or private professional use, the cost of man-hours is likely to be valued in a different way than in the office: the time required for more careful text entry with clarifications will often be spent willingly if the reward is a wider and international readership.

Eventually, in the PC and OOF world, DLT will become attractive enough not only for information distribution, but also for addressing just one (foreign) receiver. However, the path to automatically translated <u>electronic mail</u> must be paved with DLT's prior success in information distribution, and economy-of-scale benefits derived from it.

DLT is principally intended for the translation of <u>new texts</u>, when they are entered via a WP or PC. The system is NOT particularly suited for batch-wise text input from archives, neither in an OCR nor in a magnetic tapes or punched cards environment. In general, DLT is not intended for use in a central stand-alone mainframe or minicomputer installation, such as might be used by a translation service organization.

On the contrary, DLT is especially directed to the combined <u>WP/PC and network environment</u>, with all the required translation power distributed over the network and <u>built-in</u> into <u>desktop terminals</u>.

3. Product description.

We will now describe the two main DLT products, as one will be able to buy in a computer shop, presumably from 1990 on. We will indicate their practical significance from the user's point of view. Notice that in the near future the distinctions between a terminal, a PC and a videotex set will be mainly academic. The DLT products will be based on state-of-the-art (multi-processor) computers [specified in section VI.3], equipped with an optical disk unit, and built-in into compact desktop terminals. By loading them with a DLT (optical) disk, these terminals will act as translating send or receive stations; by loading them with other software, they can be turned into powerful general-purpose PC's.

i. The most simple DLT-terminal: a '<u>one-way TL set</u>'.

Let us first consider the information <u>consumer</u>, a person who uses his terminal to receive texts from various sources, including sources abroad.
If this user often accesses foreign-language sources, he may need translation into his own language. He will then go to his local computer shop and ask for a 'translating' terminal.
The dealer will make sure whether his customer only wants translation in connection with <u>reception</u> of texts, and if so, will offer him a 'DLT-terminal with a TL-module' (TL stands for 'target language', the language <u>into</u> which translation takes place). This terminal basically has the capabilities of a general purpose alphanumeric terminal, which can be used for text transmitting and receiving in the usual way without translation. The terminal must have an X25 or at least a bit-oriented (HDLC) communications interface. The translation enhancement will consist of an extra computer card (the 'DLT TL-board') and an optical disk unit. This additional hardware could take the form of a separate add-on device, connected to the basic terminal by means of a flat cable; it could also be integrated within the terminal, in one attractive housing.

The customer can specify now which TL he wants to have.
Assuming that English is his home language, the dealer will provide him with an optical disk on which the DLT-software (including a large dictionary) for the target language 'English' resides. If the same customer happens to be married with an Italian wife, he may additionally order an optical disk for Italian.
At home, this user can insert his 'English' disk and access those French, German, etc. information sources, from which he knows that they work with the DLT interface standard (presumably, this will be indicated in databank directories in the future).
Text coming from these sources will be received in English on this user's terminal now. The user will not notice any translation process: only the screen build-up is somewhat slower than for untranslated text, but the output rate (1 sentence per 2 seconds) is still high enough for him to follow with normal reading speed.
If his Italian wife uses the terminal, she will mount the 'Italian' disk, and use the same terminal set to have incoming texts translated into Italian.

Text from sources not compatible with the DLT standard can be received in the conventional way, but - evidently - untranslated. Also, all text which the user generates and sends to other terminals will be handled conventionally, without translation.

ii. The most complete DLT-terminal: a '<u>combined two-way terminal</u>'.

As a second example, let us take a person who uses his terminal to write reports for an international audience. This person can order a terminal with two translation 'directions', including a so-called SL-module (SL stands for 'source language', the language <u>from</u> which the translation takes place).
Also this terminal will have a specific DLT-board (more densely utilized than for the simple one-way terminal) and an optical disk unit. The user will normally buy just one disk, corresponding with his home language (let us again assume that this is English).
Apart from receiving foreign texts in his own language now (as described in our first example, above), the user can now also himself <u>generate</u> and transmit texts in such a way that others can receive it in their language. The procedure goes like this:

The user types his text on the terminal keyboard in the normal way, in his own language (English). After each sentence, the terminal (enhanced with its built-in translation capability) may interrupt the typing with a few questions about the last sentence. These questions concern the intended meaning of certain words or phrases, particular constructions etc. The user replies by simple selection out of several possible alternatives presented to him by the system. These question/answer pairs form a little dialogue, in which human intelligence and knowledge is called upon to assist the computer in making the correct interpretation of the sentence. The dialogues are entirely in the home language (corresponding to the DLT-disk mounted) of the DLT-terminal. On average, they will slow down the text input rate (including typing) to 1 sentence per minute.
The important result of this 'retarded' text input, however, is that the source text has now been converted to DLT-interface form, fit for transmission and reception by other DLT-terminals.

If the user prefers to enter his SL text uninterruptedly first, without being distracted by DLT's clarification dialogues, he may do so. In the queued-translation mode provided for this, he can apply the translation command to a ready, previously typed SL file. He then has to sit at his DLT-terminal and merely respond in a series of computer-initiated clarification dialogues.

In addition, the terminal can be used for conventional text input and output without translation. Again, the terminal must rely on X25 or at least an HDLC-based protocol for its basic communications interface.

4. Product range.

Fig. II-3 gives a schematic overview of the DLT product range. In addition to the simple one-way TL terminal and the complete two-way terminal, the product range includes a one-way SL device.

The one-way TL terminal translates fully automatically from the DLT interface to the TL. The built-in computer is based on a single processor (apart from several peripheral processors). The translation speed is 30 sentences/minute.

The two-way terminal translates semi-automatically from SL to the DLT interface, and fully automatically from DLT interface to TL. It is based on a built-in multi-processor computer. The rate of the semi-automatic translation is estimated at 1 sentence/minute (including typing of the SL sentence [see VI-2 for timing relations]).

The DLT interface, an intermediary between SL and TL, will be often referred to in this report as 'IL' (Intermediate Language). It is of course very important, that this interface is well defined, and adhered to by the TL- as well as the SL-related modules of DLT.

In normal operation, the two-way terminal will translate only in one direction at the same time. The user may generate text for a while, entering SL sentences and participating in the disambiguation dialogues. He may then decide to stop the entry and translation of SL text, and switch over to reception and translation of text from other sources.
The TL to which the user wants his received texts translated, and the SL in which he generates his own texts, will normally be one and the same language. A complete two-way terminal will be a package of an SL-IL and an IL-TL module, with TL=SL.

For test purposes, the two-way DLT terminal can operate in a loop mode: of each sentence generated by the user in the normal way, a so-called 'back-translation' (SL-IL-TL, with TL=SL) is displayed at the screen (this back-translation need not be exactly equal to the original SL text: slight changes of wording, including enrichments resulting from the human-assisted DLT process may occur [see V.4]).

Fig. II-3a. One-way TL terminal, permitting the reception of text in one's own language, regardless its origin.

Fig. II-3b. Two-way terminal, which also permits generation of text in one's own language, in such a way that others can receive it in their's. The dashed arc indicates the possibility of a loop-test.

Fig. II-3c. One-way SL-terminal, which only permits text generation.

Fig. II-3. Symbols for DLT products, to be used in the illustrations in this chapter (the presence of 'mirrored' symbols merely serves presentation convenience).
SL = Source Language, TL = Target Language,
IL = Intermediate Language.

As for prices, the one-way TL terminal may be favorably affected by the prospect of larger quantities in the typical information distribution environment.
Technically, the cost difference between a one-way TL and a two-way terminal (the latter having more processor and memory chips) is likely to be flattened out by the required presence of an optical disk unit in both types of terminals. Section VI-3 gives a cost range based on projected retail prices of components.
Presuming economy-of-scale effects, the end user price of a DLT terminal will come within the professional PC range. In fact, the DLT terminal is nothing else than a powerful PC, satisfying a set of minimum requirements (as to type and number of microprocessors, dynamic RAM and disc storage capacity, type of commmunications interface, etc. [see also section VI]). This hardware can be loaded with a DLT disc, but also with other software, general purpose or special purpose. The procurement of a DLT terminal will consist in the buying of a DLT (optical) disc, in addition to the required PC hardware.

The one-way SL terminal [fig. II-3c] is not so much an end-user product, but rather a building block in configurations such as fig. II-4b.

5. System configurations.

From the above product description, it will be clear that DLT is a translation system designed for use in networks and terminals, NOT for isolated stand-alone equipment.
For large and fast batch installations, offering a centralized translation service (even if the requests can be submitted on-line), other translation systems have been or are being developed (with SYSTRAN [Toma, 1977] as a well-known example).

Fig. II-4a and fig. II-5 illustrate the principal philosophy of DLT: terminals are separated by a storage and transport network, which can be thought of in abstract terms as a separation in space and time.
This separation is bridged by the DLT intermediate language (IL), the interface standard between text-generating and text-receiving terminals. Storage and transmission of textual information in a multilingual environment take place in IL, a 'semiproduct' of translation. The network simply passes this semiproduct (no translation activity at all takes place within the network). In IL-form, text may be stored and filed temporarily or permanently, inside or outside the network, just like any other kind of computer data.

Due to the difference in translation throughput between the sending and the receiving terminals, IL material needs to be accumulated and buffered at the SL-side, before a coherent piece of text can be transmitted and displayed at the TL-side at proper speed. On average, the generation of IL from SL (including SL text typing and dialogues) will last 30 times as long as the final translation from IL to TL. The rhythm and speed of SL-IL translation, IL-transmission and IL-TL translation are therefore completely decoupled.

Fig. II-4a indicates information distribution: one information provider (IP) serves consumers with different TL's. It is not necessary for the IP to individually address the consumers: one broadcast is sufficient to reach them all. Each consumer's terminal, whatever its TL, will pick up the same data packets, containing one and the same IL-version of the text. Depending on their respective TL's, the receiving terminals will then further process the IL semiproduct and convert it into different languages.
Notice that one of the TL's may equal the SL: this apparently unnecessary suit of translation and back-translation is a logical consequence of the multilingual capability or flexibility (in computer jargon one could refer to this as 'language transparency') brought about by the DLT interface. It does not harm the user in any way, if his home language is among the various different language sources which he wants to receive via DLT.

Fig. II-4b shows a battery of parallel DLT receiver terminals, displaying the same text in various languages simultaneously. Such a set-up could be useful in the editorial workshops of big IP's, who want to monitor resulting TL texts.
Another possibility is a set of parallel DLT terminals (for each language one) in public buildings, libraries, etc. Such a set could also be realized by one terminal in combination with a juke-box type of facility [MEGADOC, 1983] which permits to swap discs (and thus languages) at request.

The use of DLT for electronic mail is illustrated in fig. II-5. If the receiving terminal is unattended, it can translate the IL to TL and store the result for later inspection. Alternatively, the incoming IL stream can be stored. The latter complies more with the general DLT philosophy, according to which all internal text-handling inside multilingual computers and networks (storage, filing, transmission) is to be done in IL-form [see also section 7].

For the initial distribution and the subsequent periodic updating and maintenance of DLT software, a DLT support centre will be available. This centre supplies dealers and users with DLT with software on optical discs. New system releases, in

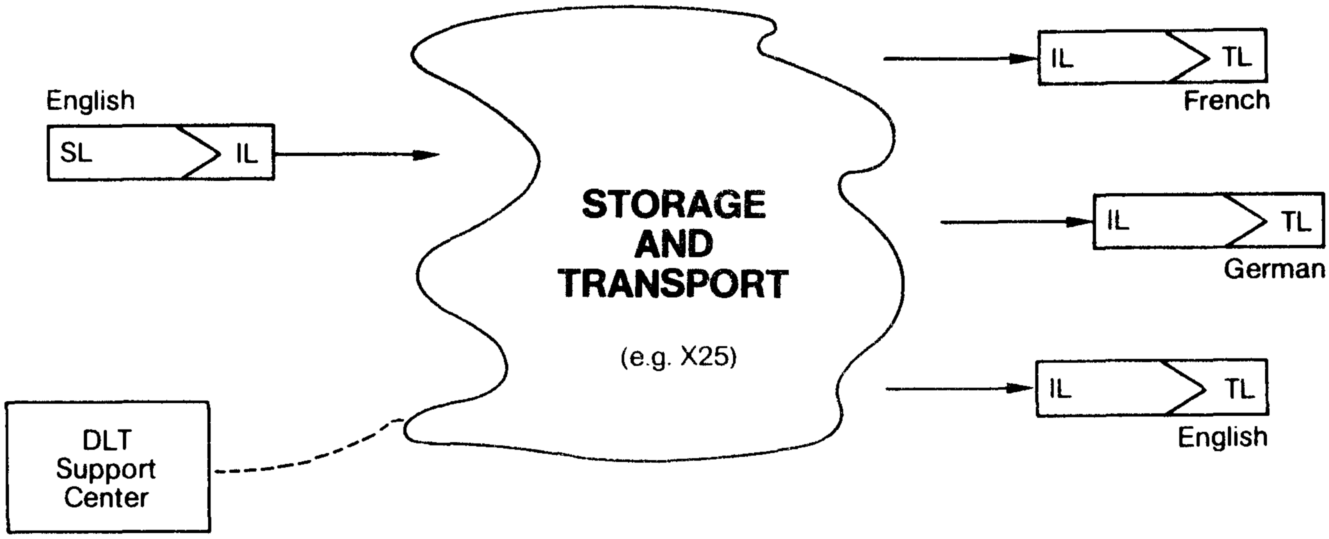


Fig. II-4a. DLT configuration for information distribution. The information provider is at the left side of the network. He enters his texts in English. The IL is used for all storage and transport, including long-term storage in a databank. The databank can be accessed by the information consumers at the right, where TL terminals convert the information to their home languages.

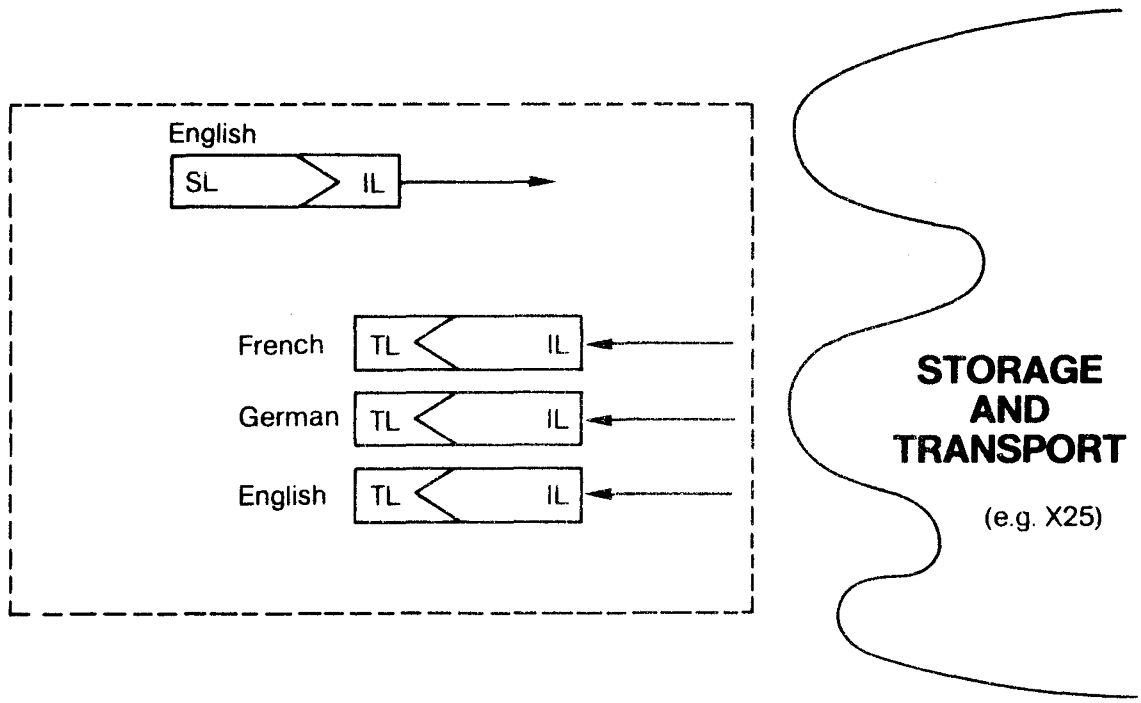


Fig. II-4b. Close-up of a big information provider's facilities. In his premises, he may monitor the various translated-text versions that his clients receive.

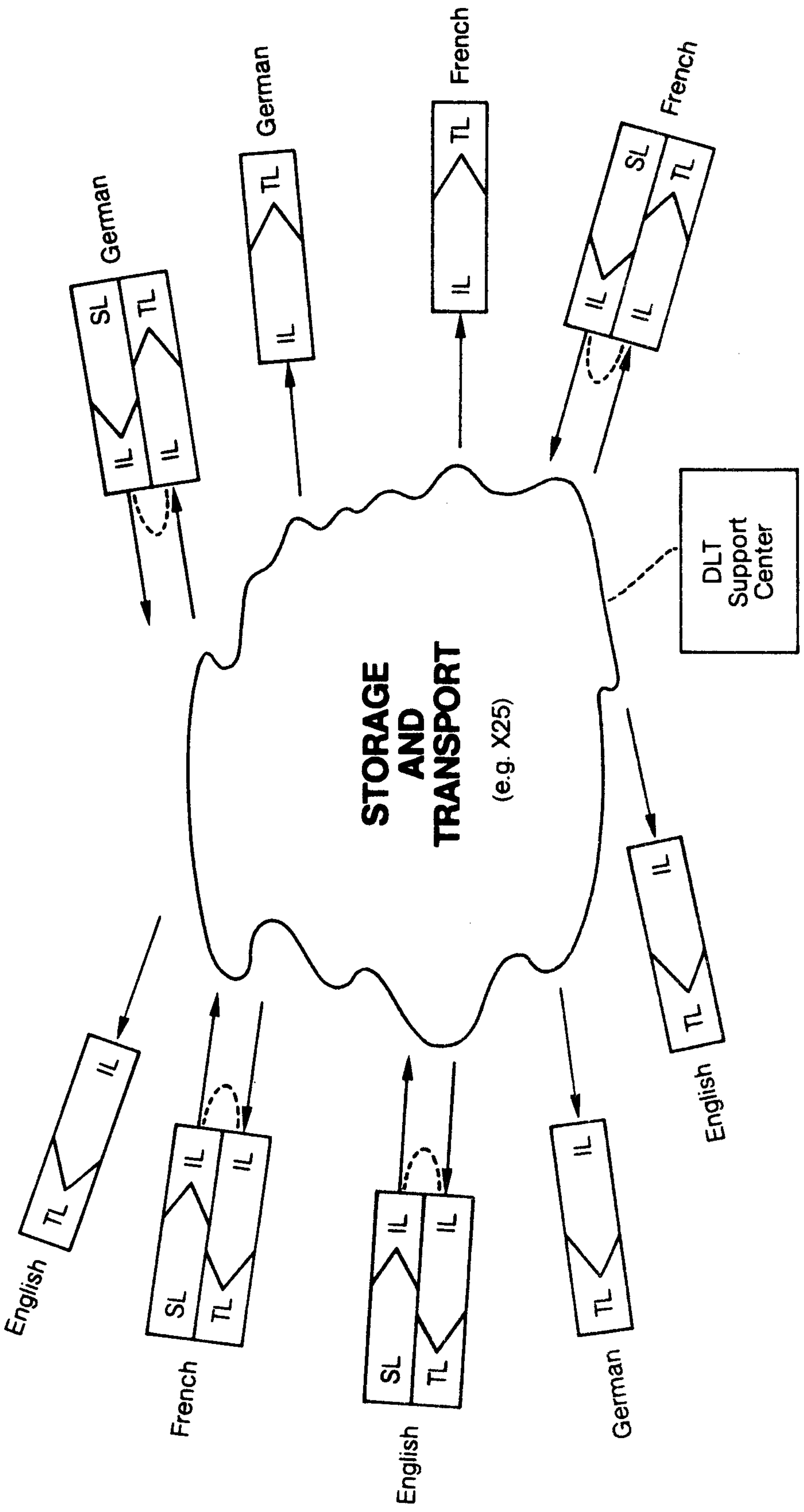


Fig. II-5. DLT configuration with several two-way stations, enabling language translation for electronic mail, in addition to information distribution.

particular periodic dictionary updates and extensions, may be arranged directly to DLT terminals by a down-line loading procedure (outside prime time).
It must be emphasized, that apart from periodic (e.g. once a month) dictionary updates, the DLT translation process relies entirely on dedicated processors and memory inside the terminals. This means that all the extensive algorithms and dictionaries required for language translation are replicated on each single terminal. Though this may seem a waste of resources at first sight, one should consider the continuing trend of decreasing storage costs compared to largely constant transmission costs. Also, the update frequency of dictionaries and other factors favor a distributed-database approach here.

6. The DLT standard.

The IL (Intermediate Language) passed through the network between sender and receiver is the pivot and backbone of DLT. On its effective standardization depends the feasibility and the expected success of the whole system.

The standard proposed for the IL interface as part of the DLT design is BCE (Binary Coded Esperanto), and will be dealt with in the body of this report [specifically in Chapter IV].

A well-defined standard for the IL is of primary importance, if one realizes that the various SL- and TL-modules (the software corresponding to distinct SL- and TL-terminals) will largely be developed by separate teams of translators and linguists. DLT products may eventually be offered by different manufacturers, e.g. one covering the consumer-electronics market with DLT receivers, another specializing in WP-like equipment for IP's. There needs to be confidence in mutual compatibility then.

7. Compatibility with existing standards.

The translation from SL to IL, and again from IL to TL can be regarded as a pair of conversions from external to internal data format vice versa.
In a multilingual text storage and transport network, several criteria exist: in addition to a well-organized distributed translation capability, there are also the costs of transmission and storage. The IL should therefore meet the typical requirement of an internal data format, i.e. compactness [see section IV.5 for details].
The code conversion from and to external format logically fits into layer VI of the ISO's OSI (Open Systems Interconnection)

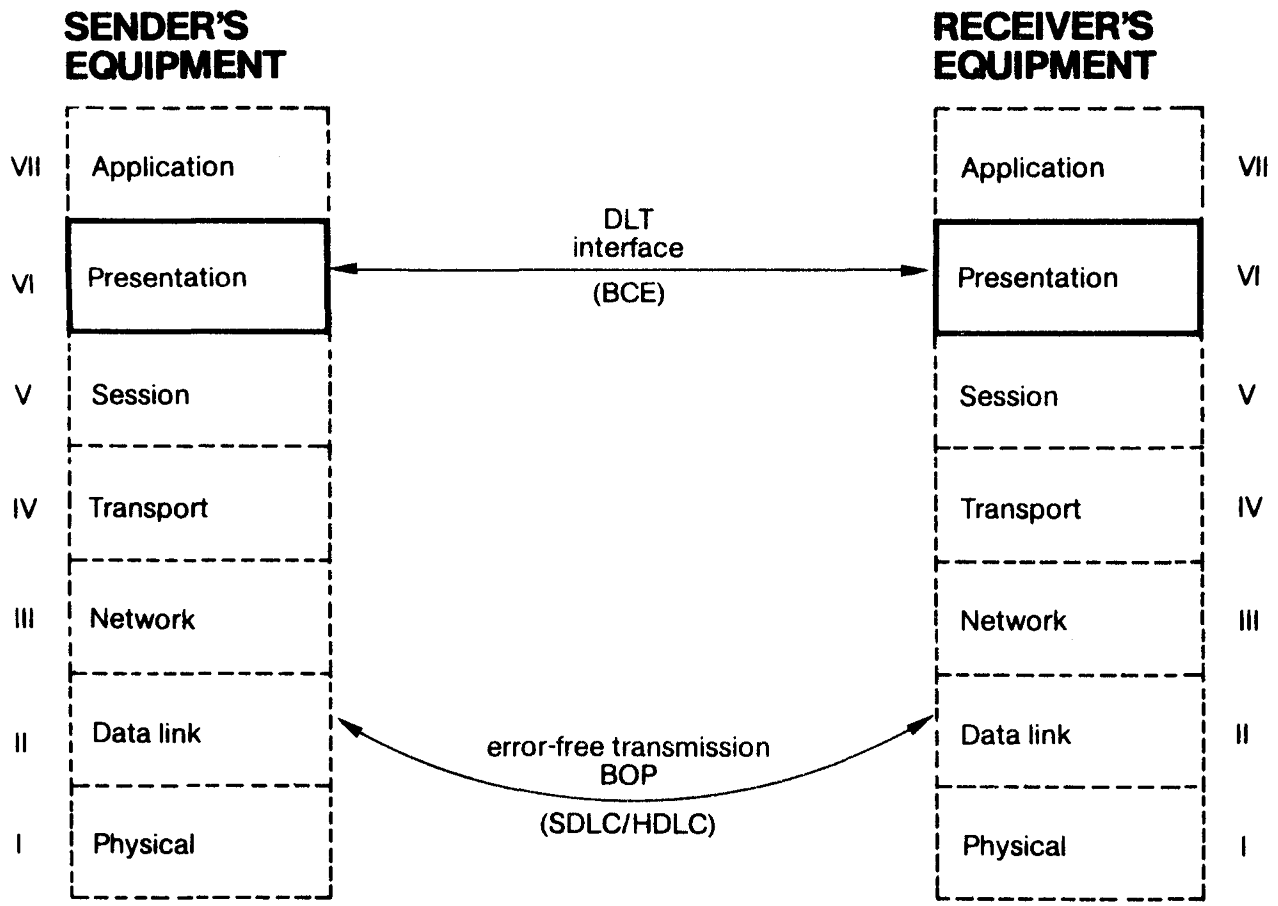


Fig. II-6. Relation of DLT with the ISO's Open Systems Architecture (OSI) reference model.
At the Presentation level (Layer VI), DLT terminals will cooperate via BCE (Binary Coded Esperanto), the standard which DLT has adopted for its intermediate language (IL).
DLT further depends on error-free transmission by a bit-oriented protocol (BOP) at the Data Link level (Layer II).

reference model [see fig. II-6]. The display of a received IL-file in the one or the other language (TL) could be regarded as a presentation choice inside the local terminal.

Connected with the presentation of different languages is the problem of foreign characters. For languages with Latin-based character sets, DLT terminals will be made compatible with the ISO 2022 standard (a shift-in extension of ASCII, formerly known as POLYGLOT-C).

As the DLT language translation is a function of the OSI's Presentation Layer, the internal BCE format (the IL) can be handled transparently by the lower layers of the OSI model, i.e. without regard to the meaning of its contents. One condition must however be met: a bit-oriented protocol (BOP) such as SDLC is required at the Data Link layer. This is because BCE is NOT character-oriented, and because transmission integrity is vital for DLT: an uncorrected bit error would not change a single character (such as sometimes occurs in ASCII-based uncorrected videotex transmission), but could distort a transmitted text entirely.

8. Scope and future of DLT.

DLT is primarily a system for translation between natural languages. The scope of languages that can be covered is theoretically unlimited. The particular architecture and choice of IL [see Chapter III] permit a modular development and steady addition of new language modules.
Some languages will be inherently more difficult (i.e. more costly) to add than others. However, a backlog of previous research and MT experiments as well as the availability of proper staff will be a more important factor.
Together with market prospect considerations, this means that French, German, English will be among the languages covered first. Other European languages like Italian, Spanish, Dutch and Swedish might follow soon.
Given an economic justification, more difficult languages like Japanese or Arabic could be taken on at a later stage. The need for different character displays is only a superficial complication compared to the development of the translation software for these languages. One must realize however, that one pair of modules (e.g. IL-Japanese and Japanese-IL) will suffice to integrate a new language into the system and make it accessible from all other languages already supported.

As has already been stated, DLT is limited to written texts. Moreover, DLT is a system to be applied at texts when they are entered via WP-like equipment, i.e. a system for new texts rather than archival texts. In connection with this, an important spin-off of DLT is that it performs a highly sophisticated spelling check ('on-the-fly', not only lexical but also grammatical, with automatic spelling correction in certain cases).

The type of text that can be handled with DLT is NOT narrowly limited. There are certainly constraints: colloquialisms, vogue idioms, slang, word games and equivocalities cannot be translated with it. The same is true for very sensitive texts, such as diplomatic writing with a high content of 'reading between the lines'. This also rules out esthetic materials and advertisement.
Apart from these obvious restrictions, the scope of text types fit for DLT ranges from 'news for the deaf' to scientific articles, and from minutes of meetings to maintenance manuals [see also Chapter V].

Beside translation from one natural language to another, the DLT architecture offers a very promising basis for the development of multilingual natural-language DB front-ends, control interfaces etc. The underlying IL appears to be an ideal intermediary between natural languages on the one hand, and formal (programming) languages on the other [see also Münnich, 1976]. Also, DLT's IL offers the best prospects for long-term maintenance of a multilingual system.

Summarizing, DLT products can be expected to play an important part in local and worldwide communications in the next decade. The terminal units proposed will then be affordable by private persons as well as companies with a need for text translation. Eventually, DLT will also provide natural-language access to to computers in various professional and consumer-oriented applications.