NATURAL LANGUAGE TRANSLATION IN INFORMATION DISTRIBUTION NETWORKS[x]

A.P.M. Witkam

BSO/Automation Technology b.v.
P.O. Box 8348 3503 RH Utrecht
The Netherlands

SUMMARY

Natural language translation by computers, traditionally concentrated at large batch installations, has come into a new environment, marked by word processors, intelligent terminals and large-scale information distribution through networks. A new approach is proposed, commensurate to a multilingual user population, an open systems architecture and economic transmission. In this approach, the translation process is distributed in space and time over the text originating and the text receiving equipment. An intermediate language, encoders and decoders are required for realization. The characteristics and selection of these components are discussed and compared against conventional techniques.

1. INTRODUCTION

Attempts to translate natural language by a computer have been undertaken since the 1950's. Also known as Machine Translation (MT), this specialized field enjoyed a period of great optimism till about 1964, when Bar-Hillel [1], one of the experts, made a notorious statement saying that "Fully Automatic High Quality Translation" (FAHQT) would be impossible. Together with other circumstances, such as the ample supply of human translators and a decreased interest for Russian scientific literature in the USA in the 1960's, this introduced a period of relative inactivity in the field. In a recent survey, Bruderer [2] lists 50-60 MT-systems around the world, most of them parts of research projects at universities or institutes.

All these MT-systems work in a typical batch environment, on large mainframe computers (CDC, IBM).with massive memories. Some 5 systems are productive on a commercial scale. One of these, SYSTRAN, has been scheduled to provide service to EURONET-users from 1982. In accordance with Bar-Hi1lel's ban on FAHQT, all known MT-systems suffer from at least one of the following limitations:

- limited field of application (e.g. only weather forecasts);
- limited to certain sentence structures (requiring human pre-editing);
- limited quality of output (requiring human post-editing);
- unability to handle ambiguities (requiring human interactive assistance).

x) First published at the IFIP/UNESCO computer network conference COMNET '81, 11 - 15 May, Budapest, Hungary.

At the other end of the scale, electronic pocket dictionaries have appeared on the market. Apart from being gadgets, these consumer products may popularize the idea that chips can store and process language, and thereby indirectly favor the acceptance of computerized translation in general.
During the last few years, a renewed interest in computer-aided translation can certainly be reported from big institutional users like the European Community (EC), whose translation load increases geometrically with its membership. As may at first glance appear natural, this interest seems to be more directed towards the batch-mainframe than towards the chip-consumer-products approach.

However, the conventional batch-mainframe approach to MT does not fit very well in today's information society. It has been based upon the classical view of language (fig. 1), which is limited to the

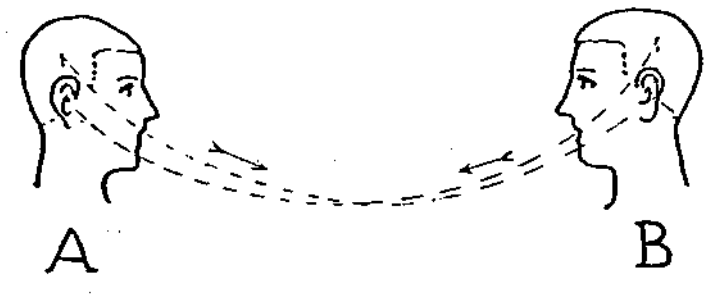


Fig. 1 Classical view of language as a 'live' communications circuit.

simple communication circuit between sender and receiver (who may or may not be capable of change roles). In that approach, the activities of the sender himself, the way in which he prepares and generates information, are clearly outside the scope (or 'system boundaries') of MT. This is perfectly illustrated by the main application of early MT in the USA: the scanning of Russian science texts.

In the age of prepackaged information, entered by word processors, stored, transported and distributed as a commodity, persistence to the classical view tends to unnecessarily isolate MT from its changed surroundings and leads to sub-optimization (fig. 2).

According, to the conventional approach, computerized language translation is just another black box in the chain, labeled 'MT'

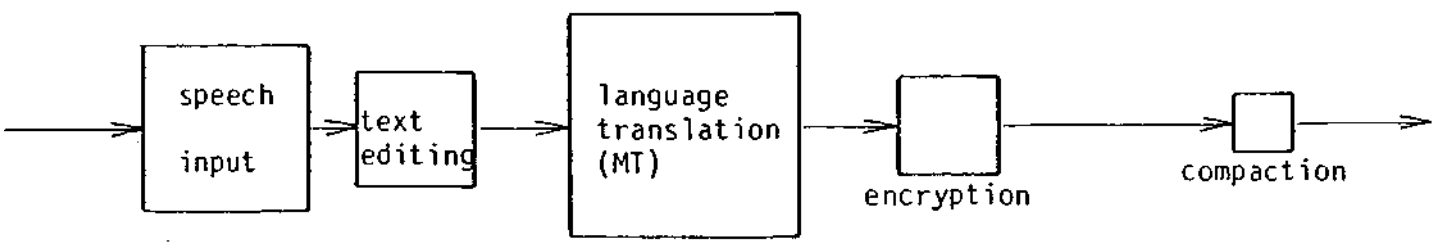


Fig. 2 Separate optimization of each black box may not result in an optimal chain.

and left to the specialism of computer linguistics. Unfortunately, that black box appears to be very ponderous and expensive, a kind of big humming monstrum in the cellars of a large organization's headquarters, and without prospects of becoming perfect at all in the near future.

2. THE NEW ENVIRONMENT

A more pragmatic and fruitful approach to language translation requires a re-orientation and appreciation of the changing environment in which it is likely to operate.

This environment is characterized by the dominancy of networks, from the so-called local networks intended for the paperless office-of-the-future to the new public and international services like VIDEOTEX and EURONET-DIANE. In this context, we use the term 'network' in a comprehensive sense: we assume it to include not only the transport function, but also the storage function, i.e. the databases themselves and the more and more popular 'mailbox' facilities.

The possibility and use of intermediate storage in the communications circuit represents one important difference with the classical 'live' circuit, and at the same time an advantage over it (2/3 of conventional telephone calls fail). Another important difference lies in the sender-receiver relationship: instead of a 1-to-1 process in classical examples, "words directed to the attention of one person have become rare" (Illich). In the office environment, only few memo's are limited to one receiver. In public networks like VIDEOTEX, receivers outnumber senders and it is more enlightening to speak of information consumers and producers. Technically, consumers have far less capabilities of generating and sending than producers. The 1-to-many traffic is often one-way.

Thus, in a more up-to-date view, language is part of the non-live and unpersonal distribution of information via a network (fig. 3). As the network now stands in between the language originator and user, the predominant operations are adding information to it (text generation)and tapping information from it (text presentation).

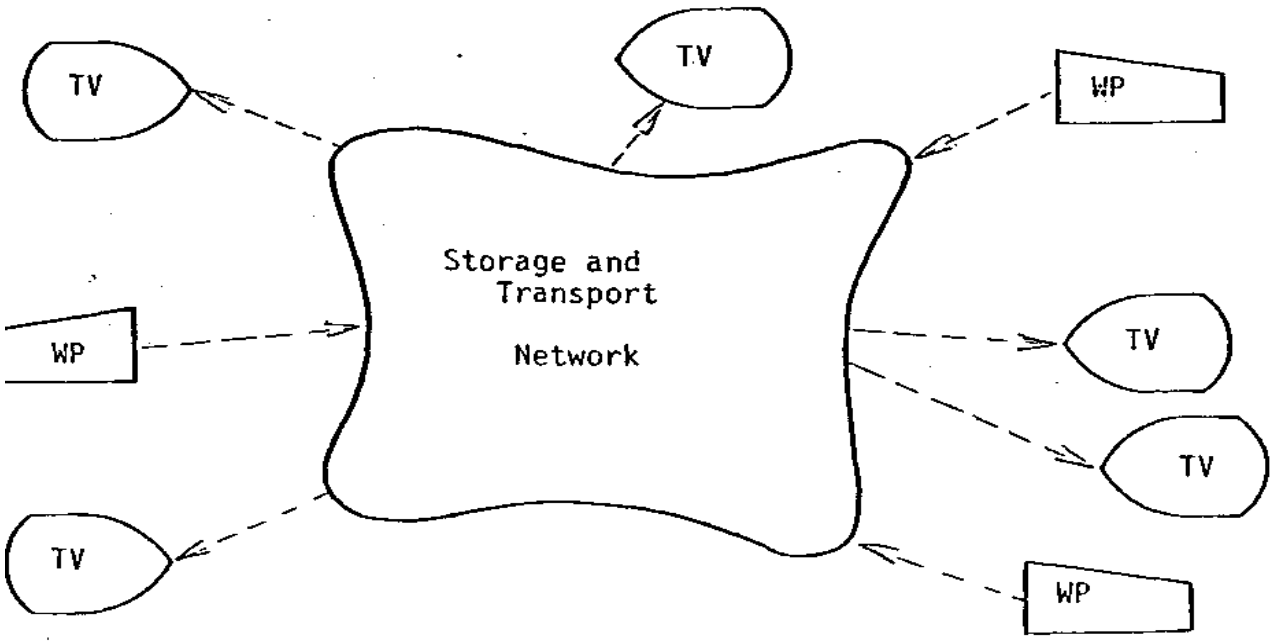


Fig 3 New view of language as part of the information on flow and out of a network. Text generators and presentators symbolized by WP (word processor) and TV (television) respectively.

These two operations are fundamental because they coincide with the conversions between external and internal format vice versa. When data or texts reside in databanks or travel through transmission lines, their format is or should be determined by the constraints and requirements of equipment and media (compaction, robustness). Only as soon as the information comes above the surface, i.e. gets visible on a terminal screen or audible from a synthesizer-loudspeaker combination, the human interface requirements become evident.
Specifically, the 'surface' is the bottom of the Presentation Layer of ISO's Open Systems Interconnection (OSI) Architecture , widely known now. The need for compaction and security on the one (internal) side, and the desire for a variety of options (voice I/O, graphics, colour) at the other (external) side, present interesting prospects for firmware-implemented conversion functions at this interface. The growing availability of powerful memory and processor chips makes new and creative solutions possible here.

3. DISTRIBUTED LANGUAGE TRANSLATION (DLT)

The question to be answered now is: Where does language translation fit? In the classical model (fig. 1), there is only one place where it can possibly be: right in between the sender and the receiver. In the new environment (fig. 3), three instead of two parties can be discerned (producers, consumers and the network), which leaves at least two possibilities: language translation concentrated between network and producers or between network and consumers.

But apart from solutions in which the whole language translation process is concentrated at one point (with disadvantages and limitations as described in a previously published paper [3]), a far better use of the new environment and its OSI-based separation of internal and external information format is made by splitting the language translation process in two parts (fig. 4). One part then becomes the conversion from readable text in (say) English to an internal format (which need not be read by anyone), the other part will be a reverse conversion from this internal format to a readable text again, but not necessarily in the same language.

Essentially, the choice of language in which information is to be presented should be a local affair of the receiving station, in the same way as other options (audio or video, character size, font etc.) that depend only on the arrangements at the user's receiving equipment. In other words, the choice in which language information will be presented, should not already be predetermined by the sender or the network. The intermediate stage during which the information is stored and transmitted in internal format, should leave that choice open to the receiver. Only under this condition, the ISO's Open Systems Architecture will be open with respect to language of presentation as well.

Being associated with an intermediate stage, the internal format can be called intermediate format. When dealing with textual information, this format is referred to as intermediate language (IL), and the conversions between natural languages and the IL can then be called IL-encoding and IL-decoding (fig. 4).

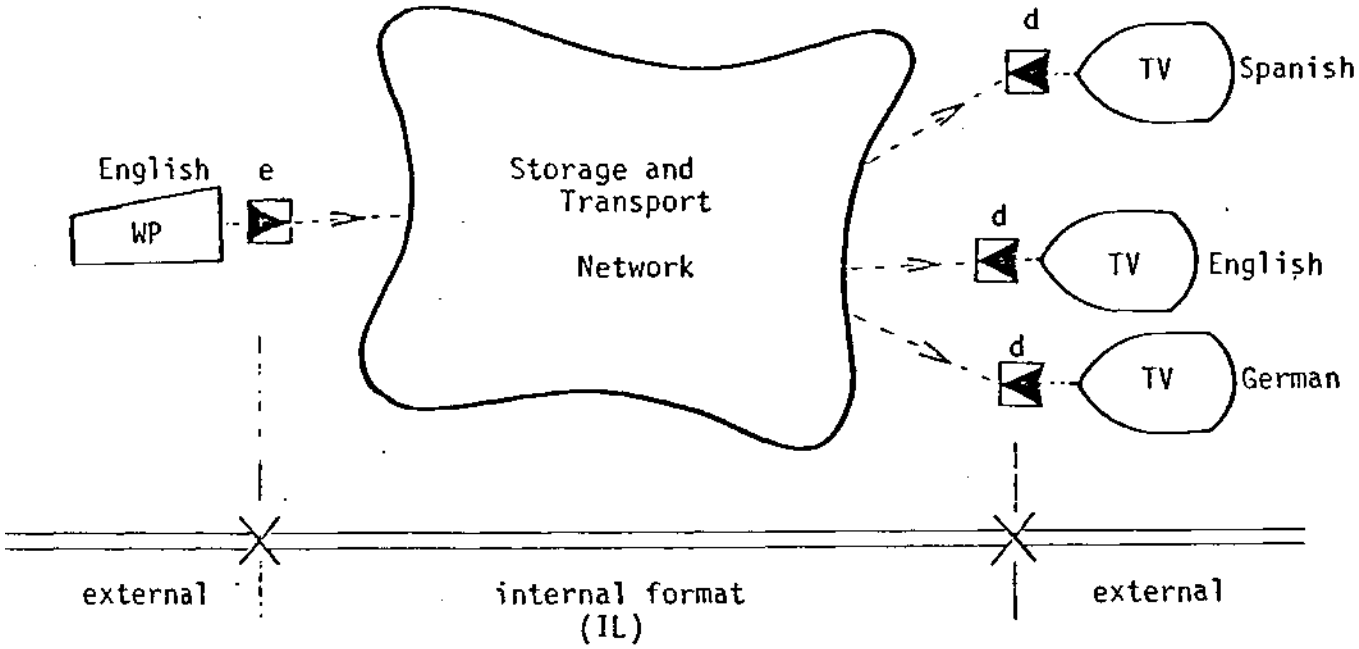


Fig. 4 The language translation process distributed over two external-internal conversions: IL-encoding (e) at the generator's and IL-decoding (d) at the presentator's site.

The concept of intermediate language (IL) has been introduced here as an abstraction arising from systems considerations about the new network environment (OSI). However, the concept has already been known throughout the history of MT.
Traditionally, the IL is the intermediate stage between the analysis of the source language and the synthesis of a target language equivalent (fig. 5a). Most MT-processes are split in these two phases, with the IL as an interface. This meant savings for those MT-centers that work with more than one language pair (the well-known m+n vs. mxn advantage), but unfortunately the interface was always restricted to the reach of just one center.

As already mentioned, MT-systems typically have been batch systems and most of them have their origin long before the era of computer networking. There seems to have been little incentive for cooperation, and each institute designed its own IL, largely independently of the others. There was no question of transmitting information in IL. As opposed to the Open Systems Architecture we considered above, the traditional IL-concept is limited by the existence of a number of closed MT-systems.

One of the characteristics of the new approach to language translation, DLT (Distributed Language Translation), is the spatial as well as time-sequential distribution of the two phases (analysis and synthesis) by transport and storage of the IL, which now becomes an interface of global importance (see also fig. 5).

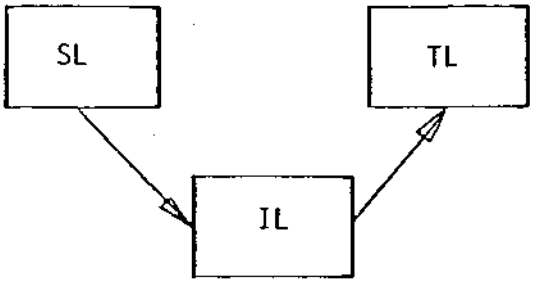


Fig. 5a Principle of IL as intermediate stage in a MT-process (IL = intermediate language, SL = source language, TL = target language).

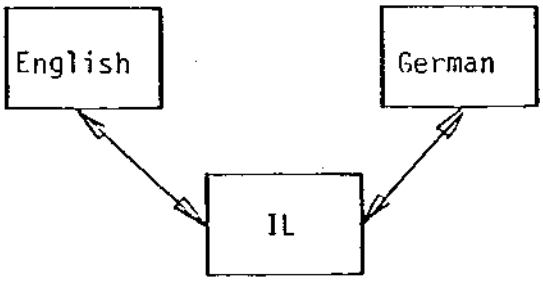


Fig. 5b Bidirectional MT-capability for a given language pair; the SL- and TL-roles change continually.

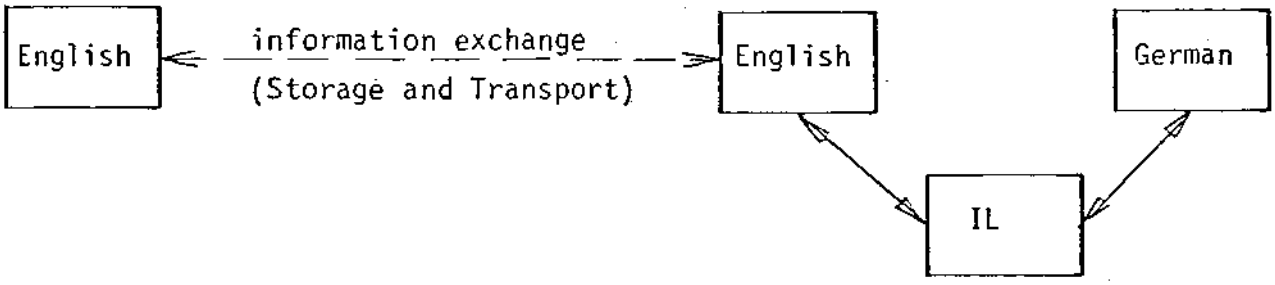


Fig. 5c A conventional MT-system. When remotely accessed, the information exchange takes place at the SL/TL-level.

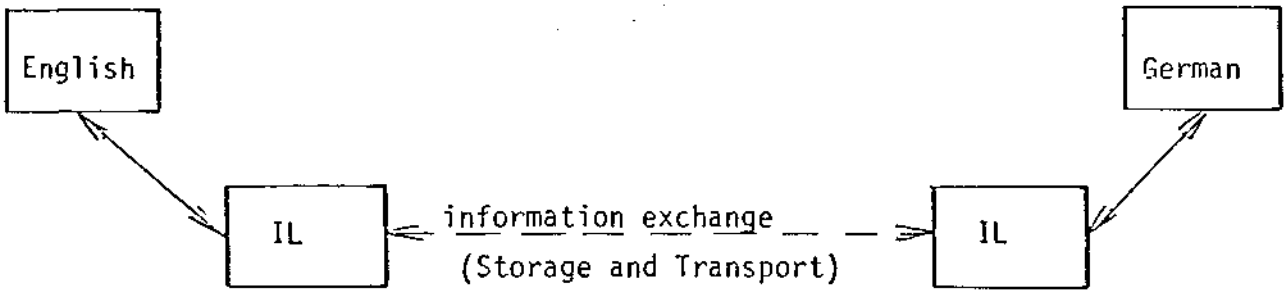


Fig. 5d A distributed language translation (DLT-)system. For remote access, the information exchange takes place at IL-level.

4. SELECTION OF AN INTERMEDIATE LANGUAGE (IL)

The decision to make use of the IL-principle is one thing, the design and selection of an appropriate IL is another. The requirements of the new environment result in the following selection criteria:

a. Standardization prospects

In DLT, the IL will provide a common interface for text generating (sending) and text presenting (receiving) equipment. These two classes of equipment may be manufactured by slightly different industry segments. Also the development of the various SL-analysis and TL-synthesis modules will be carried out by separate teams. Therefore, adherence to one standard IL is vital for the success of DLT.

b. Morphological regularity

Simplicity and regularity of the IL will enhance not only the development of the various SL- and TL-modules, but also their operational performance. The less the number of allomorphs and different paradigms, the greater the efficiency of the translation modules in terms of speed and memory requirements.

c. Compaction

Steadily increasing transmission costs and a transition towards volume-sensitive tariffs in networks form a continuous economic driving force behind compaction. Also, as the transmission rate often is the speed bottle-neck, compaction will improve frame build-up time at videotex receivers. Finally, massive storage requirements for databanks will be affected favorably.

d. Minimal ambiguity

The IL itself should be marked by the absence or a minimum of lexical and structural ambiguities. To be used as a reliable vehicle for information exchange, it should not introduce divergence in the subsequent translation phase. The requirement for disambiguity tends to be in conflict with the requirement for compaction.

The first of these four criteria forces us to look for existing candidates in the first place. These are:

I Ad hoc IL's from MT-centers:
- syntactic IL's;
- semantic or logic IL's;
- transformational IL's.

II Human communication languages:
- Ethnic languages:- English;
  - Russian;
  - Latin.
- Auxiliary languages:
  - Esperanto;
  - Ido;
  - Volapük.

From these candidates, Esperanto achieves the highest total score on the four selection criteria, leaving any of the others far behind. It also meets the requirements for a global IL so well, that its adoption appears to be more attractive than the creation of a new IL from scratch.

Let us briefly review the merits of Esperanto as to each of the important criteria:

Standardization

Well-defined, tried out, proven and survived for nearly a century, Esperanto is supported by literature and numerous up-to-date periodicals covering a rich diversity of subjects. Worldwide organizations, an international academy and a few university chairs provide a firm support, and dictionaries to and from many ethnic languages exist.

The Esperanto-'standard' is made up by several authoritative and generally recognized works:

- the PIV (Plena Illustrita Vortaro), the most comprehensive dictionary, monolingual and covering about 16.000 word roots, including scientific terminology [4];
- the PAG (Plena Analiza Gramatiko), a standard work reflecting 90 years of practical experience with the Esperanto grammar (covering morphology, syntax and word formation) [5];
- specialist dictionaries like the "International Business Dictionary", which includes and was developed by definitions in Esperanto [6].

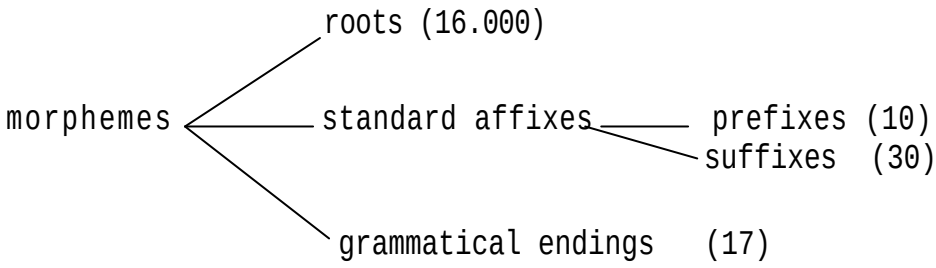
Notwithstanding a sound framework of established rules and vocabulary, Esperanto is a very open and productive language. Its word formation principles offer an enormous growth potential, and its grammar still leaves room for the creative solution of some problems that may become conspicuous in its new IL-function.

Regularity

Though at least as powerful as an ethnic language, Esperanto excels in an extremely regular and economic grammar:

- only one declination paradigm;
  absence of conjugation;
  absence of morpheme allomorphy.

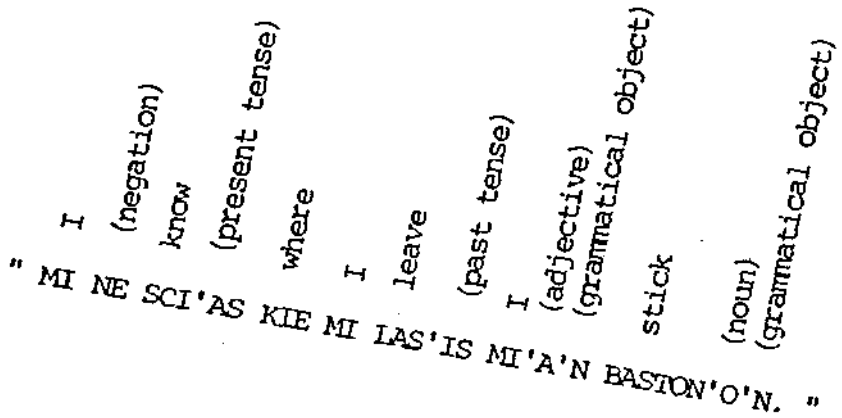
The last feature is the most inclusive and implies the other two. The unchangeableness of morphemes, both in form and in meaning, is the fundamental characteristic of Esperanto and extends over all the three morpheme classes:



The constancy of morpheme meanings under arbitrary combinations shows strong resemblance with the orthogonality principle found in the High level programming language ALGOL 68

For the use of Esperanto as an IL in computerized translation, the unique advantage connected with this morphological simplicity and stability are:

I The grammatical structure of a sentence is immediately apparent. There is no need for separate word category and syntactic function labels, as these are already included in the language itself. The following example, derived from Zamenhof [7], illustrates this (the apostrophes indicating the structure of polymorphematic words):



The sample shows the 1-to-1 relationship between morpheme and function or meaning, the existence of a distinct accusative ending and the so-called feature concord between the noun and its adjective. This enables a quick analysis of the syntactic pattern, largely independent of the word order.

II Content addressability
In the information systems discipline, this concept means that items similar in contents are stored close together in the computer's memory.
For computerized translation, this property would mean a well-structured and rational dictionary, with reduced overall access time and lower development and maintenance costs.
Esperanto meets this objective very well. Its quantity of 16.000 roots should not be compared with the much higher number of words in ethnic languages. The roots are only the basic elements with which to build words. The word formation capability of Esperanto is facilitated by an assortment of 40 standard affixes. Consequently, words similar or related in contents can be found under the same root entry in the dictionary. This applies for manual dictionaries, but can as well be exploited for computerized ones. Just compare the following samples:

SKRIB'I	write
SKRIB'IL'O	pen
SKRIB'IST'O	clerk
SUB'SKRIB'I	sign
PRI'SKRIB'I	describe

Compaction
In search for more compact coding techniques than the traditional fixed-length character codes (ASCII), word coding presents an interesting alternative. The character string associated with a word serves its external format (e.g. its presentation on a display screen), but is by no means essential for its internal representation (when stored or transmitted through networks). A word numbering scheme is feasible and brings a gain in compaction. Moreover, when language translation has to be done, a word makes at least some sense as a process element.

In Esperanto, far more than in other languages, the morpheme is the key element, thanks to its autonomy and the constancy of its form and meaning. As shown in the sample sentence on the previous page, the morpheme is the basic element in the translation process and some morphemes merely serve as a grammatical function label. Preservation of these labels in the internal format code will speed up processing of the Esperanto-IL during TL-synthesis.

The table below gives an impression of the compaction achievable by various internal format codes. As appears, Esperanto morpheme coding is not quite as compact as word coding, but the improvement over character coding is still dramatic. This is achieved by a combination of variable-length coding of the grammatical morphemes (which have a stable and peaky statistical distribution) and fixed-length coding of the lexical morphemes (the distribution of which is rather flat and context-dependent).

Thus, one can realize a compact internal format, in which the unique regular structure of Esperanto is preserved.

INTERNAL FORMAT OF NATURAL LANGUAGE TEXT	AVERAGE NUMBER OF BITS PER SENTENCE	%
character coding (English) fixed-length (ASCII)	730	100
word coding (English) variable/fixed-length	260	36
morpheme coding (Esperanto) variable/fixed-length	340	46
Andreyev's syntactic IL [8] variable/fixed-length	400	55

Table 1. The achievable degree of compaction for different internal codings. The figures are indicative and depend on several provisional assumptions (average sentence length = 18 words) and partially estimated statistical distributions. Compared to conventional ASCII character coding in languages other than English, even greater savings may be obtained.

Ambiguity

Among human communication languages, Esperanto is by far the least ambiguous. Several important types of ambiguity, widely present in other languages, are practically absent in Esperanto:

- homophony (confusion of word categories);
- ambiguity of case or number (confusion of nominative/accusative, singular/plural);
- juxtaposition problems (unclarity of the range of adjectives);

  lexical homonymy (non-related meanings of the same word).

But compared with ad hoc IL's, even Esperanto appears ambiguous. For one part, this is caused by a few peculiarities of the language (e.g. the use of certain prepositions), for the rest it is due to polysemy (related meanings and nuances of the same word).

However, by its grammar and fundamental structure Esperanto still leaves a lot of constructional freedom (word order, use of prepositions and cases). In a future IL-role, this freedom can be utilized to avoid troublesome ambiguities, by a preference for some alternative constructions and a systematic rejection of others. In this respect, [5] gives various interesting examples. Also, the rich refinement and differentiation of Esperanto's word formation offers considerable scope for counteracting polysemy.

The important thing is that with Esperanto, measures against ambiguity can be developed, verified, integrated and maintained in a language which has an excellent interface to human personnel, and which relates to ad hoc IL's as high level programming languages to assembly languages.

The ambiguity considerations when selecting an IL, as discussed here, should not be confused with the handling of ambiguities in the source language text, which will be a subject of the next section.

5. PRINCIPLES OF THE IL-ENCODER

The IL-encoding process will be arranged as an integral part of text generation on WP (word processor) equipment (fig. 6a). This is the place where a source language text enters the electronic system. It also acts like a filter, protecting the DB (databank) and the distribution network against the intrusion of rubbish and noise: whenever a problem turns up during SL-IL conversion, it is played back immediately to the human operator.

A well known principle of local autonomy nowadays is to make the end users responsible for their own data entry. Extending this to text entry, we run into the problem of source text ambiguities. The best way to tackle these is near where the text originates, i.e. by the author or his typist at the WP. His assistance will be requested in a simple computer-initiated dialogue, following the entry or submission of an ambiguous sentence. The semi-automatic analysis of the source language (SL) and its encoding into IL is thus carried out sentence by sentence, along with other tasks that require human attendance (keyboard text entry, spelling correction, text editing etc.).

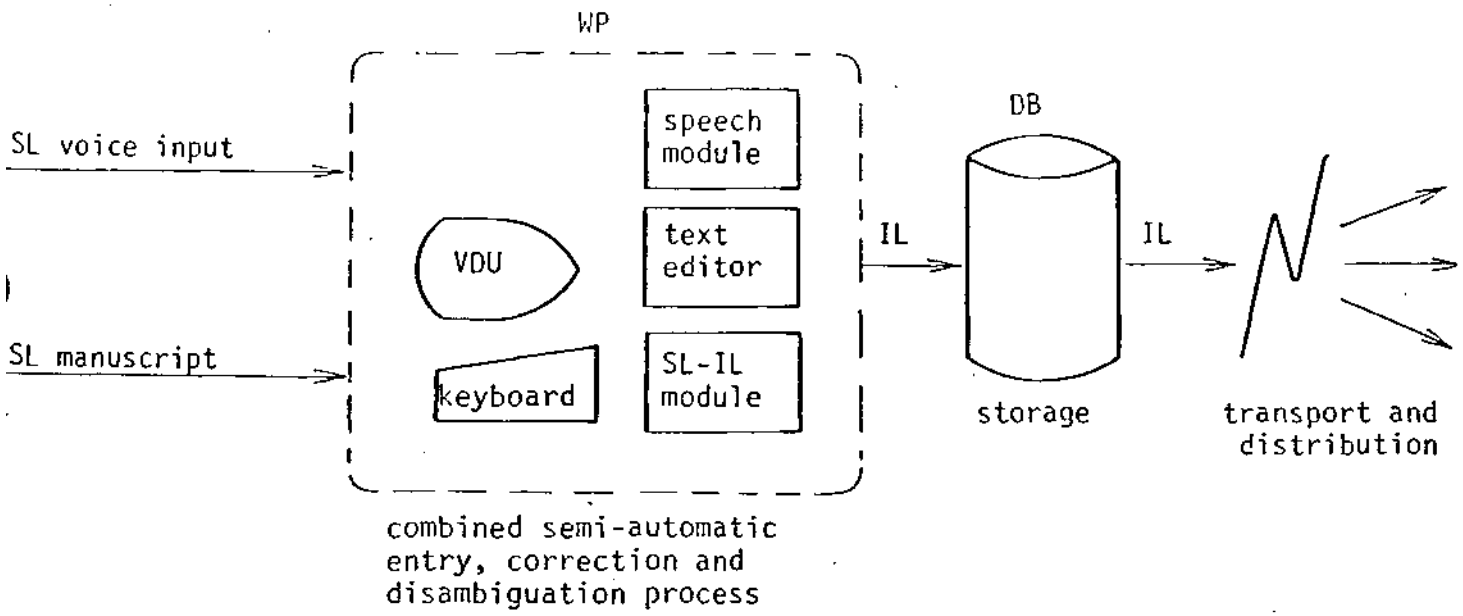


Fig. 6a Text generation in a multilingual network, incorporating conversion from source language (SL) to intermediate language (IL).

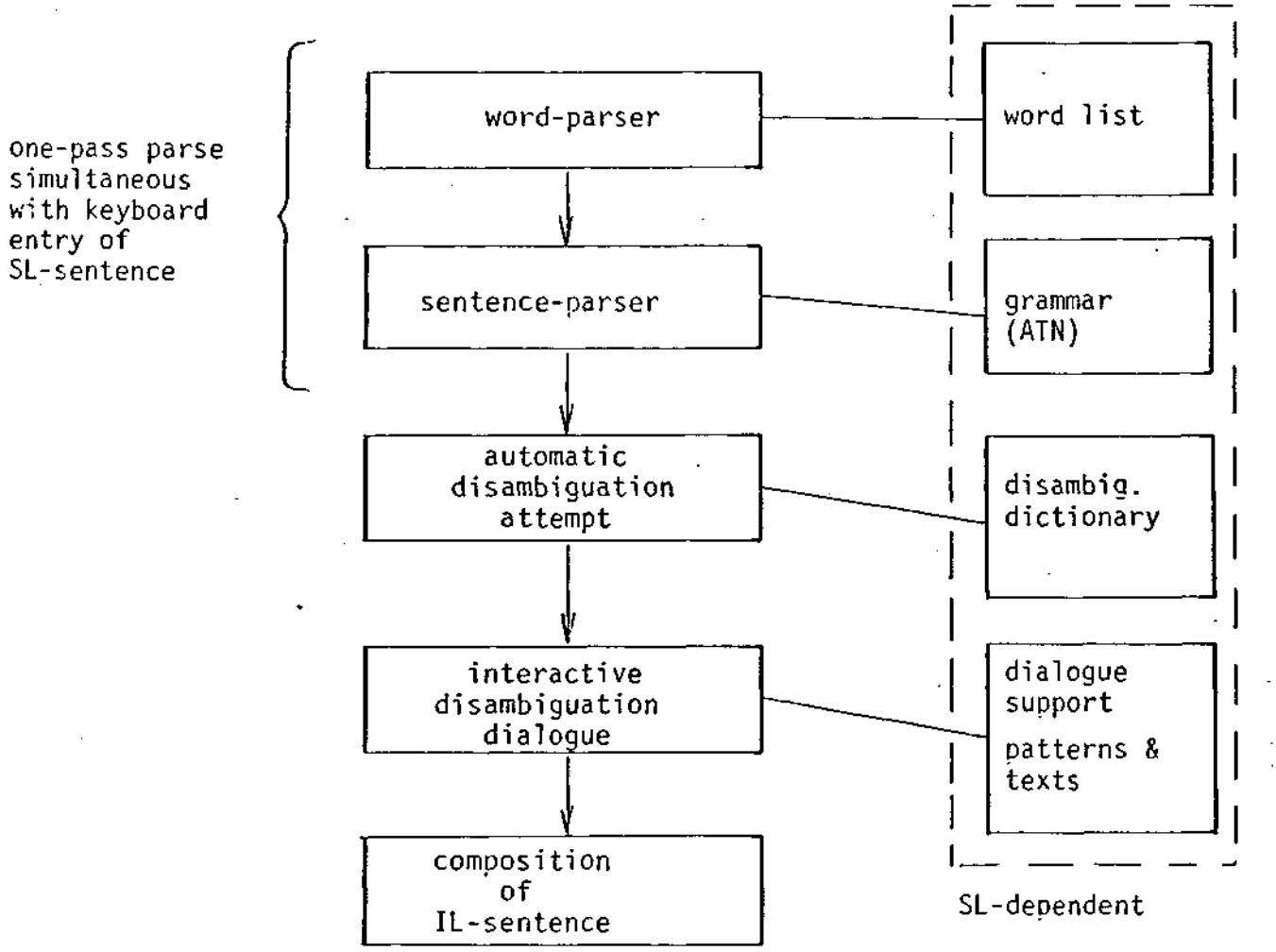


Fig. 6b Breakdown of the SL-IL conversion process per sentence. The dashed rectangle encloses the SL- dependent part of the conversion module.

Each WP-station connected to a DLT-network will be equipped with a complete set of SL-analysis programs (incl. grammar, dictionaries etc.), firmware-implemented on special-purpose microprocessor and memory boards. Like for instance a speech recognition module, such a facility will add much to the value and competitiveness of the WP. Current technology with its high-capacity memory chips will make this a viable economic product, already during the 1980's.

The major parts of the SL-IL module are shown in fig. 6b. As a sentence is being entered from the keyboard, a two-level parse is carried out simultaneously. At the word level, a stepwise access to matching word list records is obtained; at the sentence-level, one or more paths through an ATM (Augmented Transition Network, representing the SL-grammar) are traced. Because this is done during manual typing, there is no speed problem (typing speed is the bottle-neck). When otherwise (speech input) or previously entered text is submitted, a delay of a few seconds per sentence should be accounted for.

When the end of a sentence has been recognized, it is checked on unresolved ambiguities. If any, automatic disambiguation by a special built-in dictionary (micro-context approach like [9]) is attempted first. Residual ambiguities will be presented to the WP-operator, who has to decide on them by the simple choice of the pertinent paraphrase (see Table 2 below). After that, composition of an IL-sentence in internal (compact) format concludes the process.

FLYING PLANES CAN BE DANGEROUS. 1. flying can be dangerous. 2. planes can be dangerous.
THEY DON'T KNOW HOW GOOD MEAT TASTES. 1. ...how GOOD MEAT tastes. 2. ...HOW GOOD meat TASTES.
IL INSULTA LE PRESIDENT PLUS VIOLEMMENT ENCORE QUE L'ORATEUR PRECEDENT. 1. il insulta l'orateur précédent. 2. l'orateur précédent insulta le président.
SLOW NEUTRONS AND PROTONS. 1. SLOW NEUTRONS and protons. 2. slow NEUTRONS AND PROTONS.

Table 2. Disambiguation dialogue samples. From each triplet, the upper line is the sentence submitted; the other two are computer-generated paraphrases. Everything is in SL. Distinct fonts or colour will be used to highlight differences. The operator has only to type in a 1 or a 2, to indicate which interpretation applies.

In general, the text generation and disambiguation process as described above does not require bilingual or linguistic personnel at the WP. The operator should only be familiar with the source language and the context.

However, for professional information providers it may be desirable to have their end products checked by multilingual or linguistic specialists. The end products are videotex information frames in different language versions, which can be judged at receivers (with built-in IL-decoders) situated around the WP-console (this arrangement may be called a 'language organ'). Apart from variations in frame layout (due to differences in word and sentence lengths), meaning and style of the distinct language versions can be compared.

This feedback may focus attention to untranslatable idioms or metaphors of the source text. Incidentally, a translation may be improved or corrected by direct access to the IL, which then becomes the factual source. A separate pair of encoder and decoder modules, converting between internal IL and external Esperanto, will make this feasible.

6. PRINCIPLES OF THE IL-DECODER

Fully automatic, built-in into existing receivers (TV and videotex terminals), upward compatible with conventional text formats, the IL-TL conversion modules will become consumer products just as popular as already known video and synthesizer boards.

The synthesis of a sentence in the receiver's target language (TL) is a fully automatic process because it departs from an unambiguous IL-sentence. All programs and dictionaries taking part in the TL-synthesis will reside permanently on the TL-boards in the receiver. As memory components, MBM or battery backed-up RAM-chips may be employed, permitting overnight dictionary updates by telesoftware.

As to speed, the IL-TL translation takes place simultaneously with the reception of the information stream from the network. The transmission speed (1200 or 2400 bps in public videotex systems) remains the bottle-neck. DLT-users will even face a considerable improvement in frame build-up time, thanks to the reduced transmission time of the compacted IL-sentences (see Table 1). At receivers served by a high-bandwidth local network, a noticeable translation delay of several seconds per frame will accumulate.

Bad transmission line conditions however may affect the frame build-up speed. This is because the compact internal format of the IL cannot be tolerated to undergo the slightest deformation and is therefore subject to packet retransmission. A bit-oriented link protocol of the HDLC type is required in this respect, which implies the need for receivers with a synchronous or packet-switching communications interface.

The possible presence of transmission errors in incoming bit streams does not prevent the IL-decoding to proceed already during reception. Only, the presentation of any TL-results will be held back till a block or packet has passed the usual CRC-test positively.

References

1. Bar-Hillel, Y. "Language and Information", Jerusalem 1964.

2. Bruderer, H.E. "Handbuch der maschinellen und maschinenunterstützen Sprachübersetzung", Saur KG, München 1978.

3. Witkam, A.P.M. and J.J. Hillan. "Resolving language barriers in international videotex communication", Int'l Conf. on New Systems and Services in Telecommunications, Liège 1980.

4. Waringhien, G. (ed.) "Plena illustrita vortaro de Esperanto", SAT, Paris, 2nd ed., 1977.

5. Waringhien, G. and K. Kalocsay "Plena analiza gramatiko de Esperanto", UEA, Rotterdam, 4th ed., 1980.

6. Munniksma, F. (ed.) "International Business Dictionary in nine languages", Kluwer, Antwerp 1975.

7. Zamenhof, L.L. "Fundamenta Krestomatio de la Lingvo Esperanto", Paris, 5th ed., 1933.

8. Andreyev, N.D. "The intermediary language as the focal point of machine translation", in: A.D. Booth (ed.) "Machine Translation", North-Holland, Amsterdam 1967.

9. Kelly, D. and Ph. Stone. "Computer recognition of English word senses", North-Holland, Amsterdam 1975.

Patent is pending for a DLT configuration with Esperanto as IL and a variable/fixed-length morpheme coding scheme.