# Towards Personal MT : general design, dialogue structure, potential role of speech, text encoding

Christian BOITET
GETA, IMAG Institute
(UJF & CNRS)
BP 53X, 38041 Grenoble Cedex, France

## Abstract

*Personal MT (PMT)* is a new concept in *dialogue-based MT (DBMT)* which we are currently studying and prototyping in the LIDIA project, Ideally, a PMT system should run on PCs and be usable by everybody. To get his/her text translated into one or several languages, the writer would accept to cooperate with the system in order to standardize and clarify his/her document. There are many interesting aspects in the design of such a system. The paper briefly presents some of them (HyperText, distributed architecture, guided language, hybrid transfer/interlingua, the goes on to study in more detail the structure of the dialogue with the writer and the place of speech synthesis [1].

## Keywords

Personal Machine Translation, dialogue-based Machine Translation, Man-Machine Dialogue, Ambiguity Resolution, Speech Synthesis.

## Introduction

A first classification of MAT (Machine Aided Translation) systems is by user. "Classical" MAT systems are for the *watcher*, for the *revisor* (post-editor), or for the *translator*. A new concept is that of "personal MT", or MAT for the *writer*.

MT for the watcher appeared in the sixties. Its purpose is to provide informative rough translations of large amounts of unrestricted texts for the end user.

MT for the revisor appeared in the seventies. It aims at producing raw translations good enough to be revised by professionals in a cost-effective way. This implies that the system needs to be specialized for a certain sublanguage. For a system to be cost-effective, it is generally agreed that at least 20000 pages must be handled (e.g. 10000 pages/year for at least 2 years).

Leaving "heavy MT", not adapted to small volumes of heterogeneous texts, several firms have developed MAT systems for translators, in the form of tools (e.g. Mercury-Termex™), or of integrated environments (e.g. Alps TSS™).

The concept of MT for the author (writer/speaker) has recently crystallized, building on previous studies on interactive MT, text critiquing and dialog structures [5, 6, 7, 9, 12]. Its aim is to provide high quality translation/interpretation services to end users with no knowledge of the target languages or linguistics.

A second classification of MAT systems is by the types of knowledge felt to be central to their functioning. *Linguistic Based MT* uses :

- core knowledge about the language ;
- specific knowledge about the corpus (domain, typology) ;
- intrinsic semantics (a term coined by J.P. Desclés to cover all information formally marked in a natural language, but which refers to its interpretation, such as semantic features or relations : concreteness, location, cause, instrument... ) ;

but not :

- extrinsic semantics (static knowledge describing the domain(s) of the text, e.g. in terms of facts and rules) ;
- situational semantics (describing the dynamic situations and their actors) ;
- pragmatics (overt or covert intentions in the communicative context).

*Knowledge-Based MT* uses extralinguistic knowledge on top of linguistic knowledge. Finally, *Dialogue-Based MT* insists on extracting knowledge from a human (the author or a specialist). These options are not exclusive, however. In KBMT-89 [7], for example, ambiguities persisting after using linguistic and extralinguistic knowledge are solved through a dialogue with the writer initiated by the "augmentor". In ATR's Machine Interpretation project, the dialogues center around a well-defined task (organization of international conferences), but may also concern extraneous matters (cultural events, health problems...). This feature, added to the enormous ambiguity inherent in speech input, will likely force such systems to be dialogue-based as well as knowledge-based [5].

In Personal MT, we may rely on some core extralinguistic knowledge base, but not on any detailed expertise, because the domains and types of text should be unrestricted. Hence, Personal MT must be primarily dialogue-based.

A third classification of MAT systems is by their internal organization (direct/transfer/interlingua, use of classical or specialized languages, procedurality / declarativeness...) through which so-called "generations" have been distinguished. This level of detail will not be too relevant in this paper.

# I. A project in Personal MT

## 1. Goals

LIDIA (Large Internationalization of the Documents by Interacting with their Authors) aims at studying the theoretical and methodological issues of the PMT approach, to be experimented on by first building a small prototype, and more generally at promoting this concept within the MT community.

We are trying to develop an architecture which would be suitable for very large applications, to be upscaled later with industrial partners if results are promising enough. For example, we don't intend to incorporate more than a few hundred or thousand words in the prototype's (LIDIA-1) dictionaries, although we try to develop robust indexing schemes and to implement the lexical data base in a way which would allow supporting on the order of 1 to 10Mwords in 10 languages. The same goes for the grammars.

Even in a prototype, however, the structure of the dialogue with the author must be studied with care, and offers interesting possibilities. Clearly, the writer should be allowed to write freely, and to decide for himself when and on which part of his document to start any kind of interaction. But changes in the text should be controlled so that not all changes would force the system to start the interaction anew.

From a linguistic point of view, it could be quite exciting to experiment with Zemb's theme/rheme/pheme "statutory" articulation of propositions [13], and/or Prague's topic/focus opposition, which are claimed to be of utmost importance for translation : both are almost impossible to compute automatically, because the tests are very often expressed in terms of possible transformations in a given discourse context. But, in PMT, we may ask the author.

## 2. Outline

The prototype system for LIDIA-1 is constrained as follows.

- Translation from French into Russian, German and English (inversing previous systems), with other target languages being studied in cooperative frameworks ;
- Small corpus from the Ariane-G5 user interface (containing some on-line documentation), in HyperCard form ;
- Distributed computer architecture : writer workstation on a MacIntosh (Plus or SE), MT server on a mini (IBM-4361) ;
- Guided Language approach, as opposed to Free Text or Controlled Language ;
- Linguistic architecture : hybrid Transfer/Interlingua.

### *HyperText*

The choice of HyperCard reflects the fact that Hypertexts are becoming the favorite supports for technical documentation. It also relies on the assumption that writers will more readily agree to participating in a dialogue if the tool they are using is very interactive than if they use a more classical text processor. Finally, there are some linguistic advantages.

First, the textual parts are clearly isolated in fields, and not cluttered with images, formulas, tabs, markups, etc. Scripts should not be translated — if they generate messages, these must be taken from normal fields, and not directly generated (linguistic requirements may lead to better programming practices!).

Second, the textual parts may be typed, thus greatly facilitating analysis. For example, a given field may contain only titles, another only menu items, another only sentences without the initial subject (which is often contained in another field), etc. A distinct possibility is to define *microlanguages* as types of very short textual fragments (less than 2 or 3 lines, to be concrete), and to define *sublanguages* as structured collections of microlanguages for longer textual fragments.

### *Distributed architecture*

The idea to use a distributed architecture has both a practical and theoretical basis. First, we want to use the Ariane-G5 system, a comprehensive generator of MT systems developed over many years [11]. Although some micros can support this system (PC-AT/370, PS2/7437), their user-friendliness and availability are no match to those of the Mac.

Second, looking at some other experiences (Alps, Weidner), we have concluded that some parts of sophisticated natural language processing can not be performed in real time on small and cheap machines without oversimplifying the linguistic parts and degrading quality down to near uselessness. Rather, it should be possible to perform the "heavy" parts in an asynchronous but still user-friendly way, as IBM researchers have done for the Critique system [9].

Of course, this idea could be implemented on a single machine running under a multitasking operating system, if such a system were available on the most popular micros, and provided the heavy linguistic computations don't take hours.

### *Guided Language*

The "guided language approach" is a middle road between fr.e and controlled text. The key to quality in MT, as in other areas of AI, is to restrict the domain in an acceptable way.

By "controlled language", we understand a subset of natural language restricted in such a way that ambiguities disappear. That is the approach of the TITUS system : no text is accepted unless it *completely* conforms to *one* predefined sublanguage. While this technique works very well in a very restricted domain, with professionals producing the texts (technical abstracts in textile, in this case), it seems impossible to generalize it to open-ended uses involving the general public.

What seems possible is to define a collection of microlanguages or sublanguages, to associate one with each unit of translation, and to induce the writer/speaker to conform to it, or else to choose another one.

*Hybrid Transfer/Interlingua*

By "hybrid Transfer/Interlingua", we mean that the interface structures produced by analysis are multilevel structures of the source language, in the sense of Vauquois [4, 11, see also 2, 3], where some parts are universal (logico-semantic relations, semantic features, abstract time, discourse type...), while others are language-specific (morphosyntactic class, gender, number, lexical elements, syntactic functions...). In PMT, because of the necessity of lexical clarification, we should go one step further toward interlingua by relating the "word senses" of the vocabularies of all the languages considered in the system and making them independent objects in the lexical data base.

## II. Structure of the dialogue with the writer

### 1. Interactions concerning typology, terminology and style

Hence, the first interaction planned in LIDIA concerns *typology :* given a stack, the system will first construct a "shadow" file. For each textual field, it will ask its typology (microlanguage for very small texts, sublanguages for others), and attach it to the corresponding shadow record. In the case of "incomplete" texts, where for example the subject of the first sentence is to be taken from another field (as in tables containing command names and their explanations), it will ask how to construct a complete text for translation, and attach the corresponding rule to the shadow record.

The second level of interaction concerns *spelling*. Any spellchecker will do. However, it would be best to use a lemmatizer relying on the lexical database of the system, as the user must be allowed to enter new words and will expect a coherent behavior of the entire system.

Level three concerns *terminology*. The lexical database should contain thesaurus relations, indicating among other things the preferred term among a cluster of (quasi-)synonyms (e.g. plane/aircraft/ship/plane). Which term is preferred often depends on local decisions : it should be easy to change it for a particular stack, without of course duplicating the thesaurus. Note that the lexical database should contain a great variety of terms, even incorrect or dubious, whereas terminological databases are usually restricted to normalized or recommended terms. In PMT, we only want to *guide* the author : if s/he prefers to use a non standard term, that should be allowed.

Level four concerns *style*, understood in a simply quantitative way (average length of sentences, frequency of complex conjuncts/disjuncts, rare verbal forms, specific words like *dont* in French, relative frequency of nouns/articles, etc.). From the experience of CRITIQUE [9], it seems that such methods, which work in real time, may be very useful as a first step to guide towards the predetermined text types (micro- or sub-languages).

### 2. Interactions concerning syntax, semantics and pragmatics

Until now, the system has worked directly with the text as written by the author. For the remaining types of interaction, it will work on a transcription contained in the shadow record, as well as with some intermediate forms of processing stored in associated records of the shadow file. This forces to lock the original textual field (unless the author decides to change it and accepts to start again from level two).

Level five concerns the *fixed forms*. It is quite usual, especially in technical documentation, that some groups of words take a fixed meaning in certain contexts, with specific, non-compositional translations. For example, "Save as" as a menu item **Save as...** is translated in French as **Enregistrer sous...** and not as "Sauver comme", which would be correct for other uses. As a menu item, this group functions as a proper noun, not as a verbal phrase. The writer should be asked whether a given occurrence of each such group is to be treated as fixed or not. In the first case, an adequate transcription should be generated in the shadow record ("&FXD_Save_as", for example). Certain elements (such as menu items) should be automatically proposed for insertion in the list.

Level six concerns *lexical clarification*. First, polysemies are to be solved by asking the writer. For example, the word "diplôme" is not ambiguous in French. However, if translating from French into English, 2 possibilities should be given : "diplôme non terminal" ("diploma") or "diplôme terminal" ("degree"). Some polysemies are source language specific, some depend on the target languages. We want to treat them in a uniform way, by maintaining in the lexical database the collection of all "word senses" ("acceptions", not really concepts of an ontology as in KBMT-89), linked by disambiguating questions/definitions to the words/terms of the languages supported by the system.

Lexical ellipses can also be treated at that level. This problem is particularly annoying in MT. Suppose a text is about a space ship containing a "centrale électrique" ("electric plant") and a "centrale inertielle" ("inertial guidance system"). The complete form is often replaced by the elided one : "centrale". Although it is vital to disambiguate for translating correctly (by the corresponding elided forms : "plant"/"system"), no automatic solution is known. A given occurrence may be an elision or not. If yes, it is even more difficult to look for a candidate to the complete form in a hypertext than in a usual text.

At level seven, the unit of translation (the content of the shadow field) has been submitted to a first step of automatic analysis, which returns a surface structure showing ambiguities of bracketing (PP attachment, scope of coordination...). The questions to the writer should not be asked in linguistic terms. The idea is to *rephrase* the input text itself, that is, to present the alternatives in suggestive ways (on screen, or using speech synthesis – see below).

Some other ambiguities, for instance on reference (unresolved anaphora) or syntactic functions ("Which firm manages this office ?" — where is the subject ?) might be detected at this stage. They may be left for the next step to solve (actually, this is a general strategy), or solved interactively at that point. In our view, that would best be done by producing paraphrases [Zajac 1988], or by "template resolution" [6].

At level eight, the disambiguated surface structure has been submitted to the deep analysis phase, which returns a multilevel structure (decorated tree encoding several levels of linguistic interpretation, universal as well as language specific). Some ambiguities may appear during this phase, and be coded in the structure, such as ambiguities on semantic relations (deep cases), deep actualisation (time, aspect...), discourse type (a French infinitive sentence may be an order or not, for example), or theme/rheme distinction. Template or paraphrase resolution will be used to disambiguate, as no rephrasing of the text can often suffice (e.g. : "the conquest of the Barbarians").

A suggestion of [6] was to delay all interactions until transfer. The view taken here is rather to solve as soon as possible all the ambiguities which can not be solved automatically later, or only with much difficulty. For example, word sense disambiguation takes place quite early in the above scheme, and that may give class disambiguation for free.

A more flexible scheme would be to ask about word senses early only if each lemma of the considered wordform has more than one acception. If not, the system could wait until after surface analysis, which reduces almost all morphosyntactic ambiguities. A variation would be to disambiguate word senses only after surface analysis has been done. A prototype should allow experimenting with various strategies.

## III. Place and quality of speech synthesis in Personal MT

Speech synthesis has a place not only in the translation of spoken dialogues, but also in the translation of written texts. We actually think its introduction in Personal MT could be very helpful in enhancing ergonomy and allowing for more natural disambiguation strategies.

### 1. Speech synthesis and Personal MT

*Speech synthesis and MT in general*

Speech synthesis of translations may be useful for all kinds of MAT. In MT for the watcher, people could access Japanese technical and scientific textual databases, for example, through rough English MT not only over computer networks, as is currently done in Sweden [10], but also via the telephone. To produce spoken translations could be even more useful in the case of rapidly changing information (political events, weather bulletins, etc. disseminated to a large public through computer or telephone networks).

In the case of professional translation (MAT for the revisor or for the translator), the main area today is the translation of large technical documents. With the advent of widely available hypermedia techniques, these documents are starting to contain not only text and images, but also sound, used for instance to stress some important warning messages.

Personal MT could be used for translating technical documents as well as all kinds of written material not relying on creative use of language (i.e. poetry). It could also be used for communication within multilingual teams working together and linked by a network, or by phone. Finally, it could be used for the multilingual dissemination of information created on-line by a monolingual operator (sports events, fairs...) and made accessible in written form (electronic boards, minitel) as well as in spoken form (loudspeakers, radio, telephone), whence the need for speech synthesis.

Hence, *spoken output does not imply spoken input*, and should be considered for all kinds of machine aided translation. As complete linguistic structures of the translations are created during the MT process, speech synthesis should be of better quality than current text-to-speech techniques can provide. This does not apply to MAT for the translator, however (although the translator, being a specialist, could perhaps be asked to insert marks concerning prosody, rhythm and pauses, analogous with formatting markups).

*Speech synthesis of dialogue utterances*

Dialogue utterances concern the communication between the system and the user, the translation process (reformulation, clarification), and the translation system (e.g. interrogation or modification of its lexical database).

In Telephone Interpretation of dialogues, all dialogue utterances must obviously be in spoken form, the written form being made available only if the phone is coupled to a screen. In translation of written material, it could be attractive to incorporate speech synthesis in the dialogue itself, as an enhancement to its visual form, for the same ergonomic reasons as above, and because spoken alternatives might be intrinsically more suggestive than written ones in order to resolve understandability is not enough : distinguishability is a new requirement for speech synthesis.

*Other types of linguistic interactions*

In disambiguation by paraphrasing or template generation (generation of abbreviated paraphrases, as it were), questions should be generated, with their focus clearly indicated by stress and prosody. For instance :

*Is the girl or the park with a telescope ?*

In the same manner, speech quality is very important if word sense disambiguation is done orally. Since some new words or new senses of existing words may be added by the user, the disambiguation processes should apply to their definitions in the same way as they do to the texts/utterances to be translated.

All preceding remarks are of course even more valid in the case of oral input, where speech is the primary means of interaction, and the quality of the signal is reduced by the transmission channel.

## IV. Text encoding and indirect preedition

### 1. Multilingual text encoding in a universal character set for the internal representation

In classical MT, one of the first problems encountered is how to handle texts in various languages using different scripts. In the old times, computer systems offered a poor choice of character sets (such as roman without diacritics, mixed cyrillic/roman without low case), so that recourse to transcriptions was mandatory.

Nowadays, almost all computer makers offer extended character sets, and "localized" operating systems. Mac OS, for example, exists in the majority of european language, as well as in Arabic, Japanese, Chinese, etc. That solves the problem for the majority of texts, the coding schemes being such that the roman alphabet is always usable in conjunction with the specific character set of the system language : it is possible to write texts containing French embedded in Arabic, or English embedded in Chinese.

However, it is still not possible to prepare a text containing Chinese, Arabic, French and English (only Xerox' Star seems to offer that level of multilinguality), such as multilingual dictionaries. Another damaging factor is that the national character sets are not represented in the same way on different computers. For example, French ASCII is not the same on PC and on Mac. But, in the context of distributed architectures such as that of LIDIA, text representations must go back and forth between computers, preferably in a readable (and debuggable) form.

That is why we have decided to use roman transcriptions for the internal representations of texts, grammars and dictionaries. A transcription consists in a character set and a method for representing textual material possibly using other characters or marks (concerning text structure, layout, or other information). Which character set to use depends on whether portability or readability is preferred.

For MT proper, the users of Ariane have come to choose a set almost identical with PL/I character set (no low case letters, the usual punctuation marks and a few special signs). This gives total portability, at the expense of readability. For example, *'*A!2 *NOE!4L , *MAC*ALLISTER VA AUX **USA'* codes *'À Noël, MacAllister va aux USA'*. In many countries, such as Thailand, low case roman letters are replaced by local character set in bilingual terminals, so that using them destroys portability.

For the micro part of LIDIA, readability is more important, and low case roman letters are always available even in the context of non roman-based systems, so that ISO 646 (international character set, based on the Roman alphabet, with upper and low case, but without diacritics) is a better choice. With that, we might code the previous example as : *'A!2 Noe!4l, MacAllister va aux USA"* (but see below).

Diacritics are represented by "special sequences" introduced by "!" because, in technical documentation, parts are often referred by identifiers where letters and digits are mixed (BA2, C5...).

Information relative to the structure or the layout of the text is represented by tags, or "markups", in the spirit of SGML (:PARAG., :SECTION., etc.). The same mechanism is used to indicate a change of language and/or character set (some languages use more than one).

### 2. Encoding of linguistic information obtainable by (indirect) preediting

In the context of personal MT, postedition is replaced by indirect preedition. That means that the text will be enriched, normally indirectly, through interaction with the author. But experimented users may well want to do part of that preedition directly, in order to bypass lengthy dialogues. Hence, these preedition marks should be readable and understandable by non specialists.

The first type of annotation concerns the word senses. In the multilingual lexical data base of the system, we may associate a number to each sense of a word (or term), e.g. 'river.1' for "a large natural stream of water flowing into a sea" ('fleuve' in French) and 'river.2' for "a large natural stream of water flowing into another one" ('rivière' in French). But that is not very readable and precludes direct preedition. Hence, we propose to add to the sense number a fragment of the distinguishing definition, called here "semantic key", usually another word or term, getting such encodings as 'river.2=into_river' or 'wall.1=outside' (Wand in German, muro in Italian, for wall as seen from the outside), or 'pen.1=to_write', 'pen.2=enclosure', and 'pen.3=style'.

ambiguities — pauses and melody may help to delimit groups and pinpoint their dependencies, while phrasal stress may give useful indications on the theme/rheme division.

In the case of non-dialogue-based systems, there are only fixed messages, and on-line speech synthesis is not really necessary, because the acoustic codings can be precomputed. In the case of dialogue-based Machine Translation, however, an important part of the dialogue concerns variable elements, such as the translated texts or the dictionaries, where definitions or disambiguating questions could be inserted.

### *Speech in PMT : synthesis of input texts or reverse translations*

Speech synthesis of input seems to be required when producing a document in several languages, with some spoken parts. It would be strange that the source language documentation not have the spoken parts, or that the author be forced to read them aloud. In the latter case, a space problem would also arise, because speech synthesis can produce an acoustic coding (later fed to a voice synthesis chip) much more compact than any representation of the acoustic signal itself.

The concept of *reverse translation* could be very useful in PMT. The idea is to give to the author, who is presumed not to know the target language(s), some control over the translations. In human translation or interpretation, it often happens that the writer or speaker asks "what has been translated". By analogy, a PMT system should be able to translate in reverse. Technically, it would do so by starting from the deep structure of the target text, and not from the target text itself, in order not to introduce spurious ambiguities (although having both possibilities could possibly help in detecting accidental ambiguities created in the target language).

Note that speech synthesis of reverse translations might be ergonomically attractive, even if no spoken form is required for the final results (translations or input texts), because screens tend to become cluttered with too much information, and because reading the screen in detail quickly becomes tiring.

## 2. The need for very high quality speech synthesis in DBMT

It has been surprisingly difficult for researchers in speech synthesis to argue convincingly about the need for very high quality. Current text to speech systems are quite cheap and seem acceptable to laymen. Of course, it is tiring to listen to them for long periods, but in common applications, such as telephone enquiry, interactions are short, or of fixed nature (time-of-day service), in which case synthesis can proceed from prerecorded fragments.

DBMT, as envisaged above, seems to offer a context in which very high quality could and should be demanded of speech synthesis.

### *Ergonomy*

First, the writer/speaker would be in frequent interaction with the system, even if each interaction is short. The overall quality of speech synthesis depends on three factors :

- voice synthesis (production of the signal from the acoustic coding) ;
- linguistic analysis (word class recognition, decomposition into groups), for correct pronunciation of individual words, or contextual treatment (liaisons in French) ;
- pragmatic analysis (communicative intent : speech act, theme/rheme division...), for pauses, rhythm and prosody.

We will consider the first factor to be fixed, and work on the linguistic and pragmatic aspects.

Of course, certain parts of the dialogue could be prerecorded, namely the messages concerning the interaction with the system itself. However, users might rather prefer a uniform quality of speech synthesis. In that case, these messages might be stored in the same acoustic coding format as the texts produced under linguistic control.

### *Ambiguity resolution by rephrasing*

We have seen two main ways of disambiguating structural ambiguities in DBMT, namely rephrasing and paraphrasing. Rephrasing means to present the original text in different ways. Suppose we want to disambiguate the famous sentence *"He saw a girl in the park with a telescope"* by presenting the alternatives on a screen. We might get something like :

| | | | | |
|---|---|---|---|---|
| *1-* | *He saw* | *the girl* | | |
| | | *in the park* | | |
| | | *with a telescope* | | |
| *2-* | *He saw* | *the girl* | | |
| | | *in the park* | | |
| | | | *with a telescope* | |
| *3-* | *He saw* | *the girl* | | |
| | | | *in the park* | |
| | | *with a telescope* | | |
| *4-* | *He saw* | *the girl* | | |
| | | | *in the park* | |
| | | | *with a telescope* | |
| *5-* | *He saw* | *the girl* | | |
| | | | *in the park* | |
| | | | | *with a telescope* |

If the disambiguation happens orally, the spoken forms should be presented in the same register as in the original (here, affirmative), but very clearly distinguished, so that a human could reconstruct the forms above. The availability of complete linguistic structures is necessary, but not sufficient, because

For readability reasons, we revert to using upper case for the main part and low case for the annotations, so that we get 'WALL.2=outside' or 'PEN.1=to_write'. The exact separators chosen (here "." and "=") should not be fixed once and for all, but be redefinable within a transcribed text. Note here that various semantic keys can be associated to the same sense (like 'PEN.1=to_write' and 'PEN.1=instrument'). Hence, if a user directly writes 'PEN=instrument', where 'instrument' is not the default semantic key, the lexical data base should be searched to determine the (unique) sense number and insert it in the encoding.

The second type of annotation concerns grammatical information relative to words, like category (verb, noun...), number, gender, case, tense... and even syntactic function (subject, object...). The main idea here is that *annotations may concern several levels of linguistic description, but should be incomplete at each level,* so that no unfamiliar notion appears. For example, 'verb' is a concept familiar to almost every literate adult, but not 'modal verb'. At the level of functions, 'subject', 'object' and 'complement' may be taken to be familiar, but not 'attribute', 'epithete', 'head', and so on.

The third type of annotation concerns the surface tree structure. For us, a structure is called "surface" if the underlying text can be read from it in a direct way (by inorder traversal of the leaves in constituent structures or of all nodes in dependency structures), although its elements (nodes and/or edges) may contain "deep" information concerning logical predicate/argument organization or semantic relations. On the contrary, "abstract" structures don't correspond with the texts in a simple way (negations and auxiliaries may have been suppressed as nodes and represented in decorations, elided elements may have been restored, ordering may have been normalized, etc.).

To delimitate syntagmatic groups, we simply use special brackets such as (REL( ... )REL) for a relative clause, or simply (( ... )) if the syntagmatic category is not known or too arcane for general users (e.g., 'adjectival group', 'cardinal group'). As for individual words, we may add grammatical and semantic information, again provided it is commonly used. For example :

'(( *COMING&ger TO THE RIVER.1 , ))&cause HE READ&verb,past A NOVEL&noun .'

Internal encoding seems at first highly technical. However, its detailed design is influenced by the overall constraint to develop systems for the general users. Incidentally, it may prove quite difficult for linguists accustomed to make very fine-grained distinctions to prepare systems using only "rustic" information obtainable from non specialists, but it must be done for that type of approach to work.

## Conclusion

The concept of Personal MT crystallizes many ideas from previous systems and research (text-critiquing, interactive MT, dialogue-based MT, Machine Interpretation of spoken dialogues, controlled languages...). However, the perspective of interacting with the author, not required to have any knowledge of the target language(s), linguistics, or translation, puts things in an original framework.

While the development of systems of this nature poses old problems in a new way, and offers interesting new possibilities to the developers, their acceptability and usefulness will perhaps result more from their ergonomy than from their intrinsic linguistic quality, how necessary it may be.

Promotion of the National Languages is becoming quite important nowadays, but, apart of efforts to teach a few foreign languages, no technical solution has yet been proposed to help people write in their own language and communicate with other people in their own languages. Personal MT could be such a solution. We strongly hope that many researchers will take interest in this new field of MT.

## Acknowledgements

Although speech synthesis of the input or output texts had been considered for the initial design or the project, and thought to be useful in other parts, it was J.I. Tsujii who pointed to me how interesting it would be to use it in ambiguity resolution, provided we can reach the necessary quality. I am also grateful to J.Ph. Guilbaud, E. Blanc, and M. Embar for reviewing previous versions of this paper. While their help was very valuable for improving both content and form, the remaining deficiencies are of course mine.

## References

[1] BOITET Ch. (1989) *Motivations and architecture of the LIDIA project.*
Proc. MTS-89, Munich, 16—18 Aug. 1989, 5p.

[2] BOITET Ch. (1988d) *PROS and CONS of the pivot and transfer approaches in multilingual Machine Translation.*
New Directions in MT. BSO conf, Budapest, 13 p.

[3] BOITET Ch. (1988c) *Hybrid Pivots using m-structures for multilingual Transfer-Based MT Systems.* Jap. Inst. of Electr., Inf. & Comm. Eng., June 1988, NLC88-3, 17—22.

[4] BOITET Ch. (1988b) *Bernard VAUQUOIS' Contribution to the Theory and Practice of building MT Systems : a historical perspective.*
2nd Int. Conf. on Theor. & Meth. Issues in MT, CMT, CMU, Pittsburgh, June 1988, 18 p.

[5] BOITET Ch. (1988a) *Representation and Computation of Units of Translation for Machine Interpretation of Spoken Texts.*
TR-I-0035, ATR, Osaka, 41 p. Also Computers & Artificial Intelligence 8/6, 505—546, 1989.

[6] CHANDLER B., HOLDEN N., HORSFALL H., POLLARD E., McGEE WOOD M. (1987) *N-Tran Final Report.* Alvey Project, CCL/UMIST Report 87/9, Manchester, 30 p.

[7] NIRENBURG & al. (1989) *KBMT-89 Project Report.* CMT, CMU, Pittsburgh, 286 p.

[8] PECCOUD F. (1988) The Aims of the French National Project of Computer-Aided Translation. In : International Forum on Information and Documentation. **13/1**, 11—13.

[9] RICHARDSON S. D. (1985) *Enhanced Text Critiquing using a Natural Language Parser : the CRITIQUE system.* IBM Research Report RC 11332, Yorktown Heights.

[10] SIGURDSON J. & GREATEX R. (1987) *MT of on-line searches in Japanese Data Bases.* RPI, Lund Univ., 124 p.

[11] VAUQUOIS B. (1988) BERNARD VAUQUOIS et la TAO, vingt-cinq ans de Traduction Automatique, ANALECTES. BERNARD VAUQUOIS and MT, twenty-five years of MT. Ass. Champollion & GETA, Grenoble, 700 p.

[12] ZAJAC R. (1988) *Interactive Translation : a new approach.* Proc. COLING-88, Budapest, Aug. 1988.

13] ZEMB J. M. (1982) *Les occurrences phématiques, hématiques et thématiques des archilexèmes modaux.* In "La otion sémantico-logique de modalité", Univ. de Metz & linksieck, Recherches linguistiques VIII, 75–116.

-o-o-o-o-o-o-o-o-o-o-