Dialogue-Based MT and self-explaining documents as an alternative to MAHT and MT of controlled languages

Christian Boitet

GETA, Institut IMAG (UJF & CNRS) BP 53, 38041 GRENOBLE cedex 9, France
Christian.Boitet@imag.fr

Abstract

We argue that, in many situations, Dialogue-Based MT is likely to offer better solutions to translation needs than machine aids to translators or batch MT, even if controlled languages are used. Objections to DBMT have led us to introduce the new concept of “self-explaining document”, which might be used in monolingual as well as in multilingual contexts, and deeply change our way of understanding important or difficult written material.

1 Introduction

In many situations, documents such as working notes, scientific abstracts, transparencies, calls for proposals, technical documentation, etc., should be translated into several languages, but are not translated, because they are ready at the last moment, and available translators have no time to do the job, or because there are simply no translators to do the job, and of course, in all cases, because no satisfactory MT solution is available.

Our first point is that interactive Dialogue-Based MT systems (DBMT), especially of the kind we are prototyping in the LIDIA project, offer a better hope to solve the problem than machine aids for translators and “black box” MT, even if controlled languages are used.

Our second point is that the DBMT approach also leads to a new and extremely interesting possibility, that of producing all versions of a document, that is, the source document and all its translations, as “self-explaining” documents, each consisting of a normal document and its deep or (even better) multilevel disambiguated linguistic representation, augmented by a memory of the original ambiguities and of the disambiguation process.

Finally, we observe that the production of self-explaining documents might also be very useful in monolingual contexts, and perhaps lead to new ways of accessing and using documents of any kind: one could “click” on any part marked as ambiguous, and get clarifying presentations or paraphrases of it. Thus, an unrestricted self-explaining text would be less ambiguous than a text in a controlled language, which may be unambiguous for a machine but not for a human, and access to texts written in foreign languages would also be facilitated. In this way, authors’ true intentions would accompany their productions in other places, times and tongues.

2 Motivations

The idea of DBMT has been proposed and experimented with in various forms during the last 20 years [13, 15, 17, 18, 24, 29, 32, 33, 36-38]. However, it has always been taken for granted that the user should be a specialist, linguist or translator or at least a professional and that consequently the system could and should be specialized. In contrast, we think that DBMT systems should be designed for the general public and should be usable on personal computers. Consequently, the design of the user interface in general, and of the disambiguation dialogues in particular, becomes extremely important.

The main idea of our current concept is that pieces of the text under creation or modification are sent to an analyzer running in the background. If there are ambiguities, be they proper to the source language or relative to the translation into one or several target languages, questions are asked of the author, in the source language. The resulting ambiguity-free structures are then sent to transfers and generators into all target languages, producing high quality translations needing no postedition.

During the last few years, we have designed and implemented a mock-up, LIDIA-1 [3-10], to experiment with this concept of DBMT for (non-specialist) individual authors. The mock-up has now only one source language, French, and three target languages, German, English and Russian, but that is only due to the limitations in manpower. This experimentation has led us to various innovations:

- *distributed processing.* The document is created and interactively disambiguated on a middle-range Macintosh, while the various processes of MT proper are performed by a distant server.

- *application to hypertexts*. The documents are in effect hyperdocuments, in the form of HyperCard stacks. Units of translation are HyperCard textual “fields”.

- *asynchronous and non-pre-emptive processing.* Units of translation are “released” by the author, and then autonomously travel to the MT server, come back after analysis in a “multiple-multilevel-concrete” form (mmc-structure), announce the presence of ambiguities by letting a button appear next to them, react to the click by engaging in a disambiguation dialogue and then, once disambiguated, travel again autonomously to the MT server to be translated and finally to be inserted in the appropriate field in the target stack.

- *e-mail communication between component processes.* We have switched from a specialized connexion to the use of standard e-mail for all communications between the author workstations and the MT server, which can now be located anywhere in the world.

- *deeper multilevel approach.* We have added a level of “interlingual acceptions” (or word senses) to the classical lexical levels of B.Vauquois’ multilevel transfer approach (“occurrence” or wordform, “lemma” or citation form, and “lexical unit” or derivational family).

- *disambiguation strategy.* We have developed a generator of disambiguation dialogues which non-specialists can easily understand and which does not rely on too sophisticated linguistic processing, so that the disambiguator can run in real time on the author’s personal computer.

- *control by reverse translation.* On demand, the system translates back from the target uma-structures (unambiguous, multilevel, abstract), providing a feed-back through a paraphrase in the source language of the translation.

- *homogeneity of knowledge sources.* In the current state of the implementation, this concerns only the lexical knowledge: both the lexical disambiguation messages and the MT dictionaries are obtained from the same multilingual lexical data base, PARAX [2], itself implemented in HyperCard.

An interesting possibility offered by our distributed technique is to build DBMT applications by using heterogeneous components. For example, the source text could be written (in French) in Paris, sent for analysis to our server in Grenoble, disambiguated interactively in Paris, and then sent to our server to produce translations in English, German and Russian, and to a server in Japan to produce a Japanese translation. This would only require appropriate “filters” (format transducers) between intermediate structures, and agreements with server operators.

3 Self-explaining documents

In the course of our experimentation, we have (again) observed that translation introduces ambiguities which are not present in the source text. *Traduttore, traditore…* It also happens that all disambiguated analyses of a sentence produce the same translation, which is as ambiguous as the original. One example was the translation from French into Russian of the famous sentence «The man sees the girl in the park with a telescope».

Then, goes the objection, what is the use of disambiguating the source text if ambiguities reappear in the translation(s), or even worse if new ones are created? Would it not be better to try and produce translations which preserve the ambiguities, and dispense with interactive disambiguation altogether?

Unfortunately, the experience of human translation shows that ambiguities can be *exactly* preserved only in some cases, and that to do it purposefully is quite difficult and often leads to unnatural ways of expression in the translated text. It is also quite clear that the “transferable” ambiguities vary with the target language. Finally, although some texts may be intentionally ambiguous, especially in poetry and politics, we take it that the vast majority of ambiguities are not intentional, but are due to the intrinsic nature of natural languages. Of course, some authors write more clearly than others, but all authors write unambiguously in any programming language, unambiguous by construction, and ambiguously in any natural language, ambiguous by nature!

This has led us to the idea of *self-explaining documents:* if the target documents are accompanied by their (unambiguous) linguistic structure, with the indications of potentially ambiguous parts, and if the reader in the target language may obtain a clarification of unclear parts in a user-friendly way, the objection disappears. As human users are notably not very sensitive to ambiguities, however, we should find a way to warn the reader that the target text is ambiguous.

In a multilingual DBMT setting, there is a very simple solution to this task. The system simply analyzes the target text with the analyzer of the target language and gets the corresponding mmc-structures. It then runs the disambiguation dialogue on the target side in automatic “mute” mode, that is by having the system itself answer each question so that the accompanying structure is contained in the selected subset at each point and memorizes questions and answers. It is then possible to show the presence of ambiguities by any convenient means, such as by creating buttons on which the reader may click to obtain the clarification *which would have been given by the author himself, were the text to have been written in the target language!* To simplify this process, the accompanying structure should then be unambiguous and “concrete”.

Let us clarify what we call “concrete” and “abstract” linguistic structures. A “concrete” representation of a text is such that the corresponding text can be recovered from it by using a standard traversal algorithm and simple morphological and graphematical generation rules. Familiar examples are textbook constituent structures and dependency structures (with left-to-right traversal of the leaves or infix traversal of all nodes). Otherwise, we say that the representation is “abstract”. Note that the information contained in both kinds of structures (on labels and other more or less complex annotations) may be of the same linguistic “depth”: there may be “deep” concrete structures and “surface” abstract structures, in this sense, although the opposite is of course more frequent.

Take for example the sentence: *“The customers were not given their money back by the cashier but by the waiter.”* A “multilevel” head-driven concrete structure could be:

```
S[type=assertive, time=past, aspect=perfective, tense=c-past, voice=passive…]
   (NP[semrel=dest, logrel=arg2, synfunc=subj, sem=human, num=plur…]
      (Art[lex=‘the’, semrel=deict, synfunc=det, number=plur, deter=definite…]
       Noun[lex=‘customer’, synfunc=head, sem=human, number=plur…])
    aux[lex=‘be’, tense=pret, pers=3, number=plur…]
    neg[lex=‘not’]
    vrb[lex=‘give’, synfunc=head, voice=passive, tense=ppart, vbpart=‘back’…]
    NP[semrel=patient, logrel=arg1, synfunc=obj1, number=sing…]
      (adjposs[lex=‘his’, semrel=poss, synfunc=det, number=plur, deter=definite…]
       Noun[lex=‘money’, synfunc=head, number=sing…])
    vbpart[lex=‘back’]
    NP[semrel=agent, logrel=arg0, synfunc=agcomp, number=sing, neg=not-but…]
      (prep[lex=‘by’, synfunc=reg]
       art[lex=‘the’, semrel=deict, synfunc=det, number=sing, deter=definite…]
       Noun[lex=‘cashier’, synfunc=head, sem=human, number=sing, neg=not…]
       NP[semrel=id, logrel=arg0, synfunc=coord, number=sing…]
         (conj[lex=‘but’, synfunc=reg]
          prep[lex=‘by’, synfunc=reg]
          art[lex=‘the’, semrel=deict, synfunc=det, number=sing, deter=definite…]
          Noun[lex=‘waiter’, synfunc=head, sem=human, number=sing…]))
    punct[lex=‘.’])
```

Syntactic categories have been used here as main labels, with phrases (syntagmas) in capitals and preterminals in small letters. Acronyms should be self-explaining.

In an abstract structure, some lexical information would be “featurized”, and order could be normalised, leading to:

```
S[type=assertive, time=past, aspect=perfective, tense=c-past, voice=passive…]
   (vrb[lex=‘give’.’back’, synfunc=head, voice=passive, tense=c-past…]
    NP[semrel=agent, logrel=arg0, synfunc=agcomp, num=sing, neg=not-but…]
      (neg[lex=‘not’]
       Noun[lex=‘cashier’, synfunc=head, sem=human, number=sing, deter=definite…]
       NP[semrel=id, logrel=arg0, synfunc=coord, num=sing…]
         (conj[lex=‘but’, synfunc=reg]
          Noun[lex=‘waiter’, synfunc=head, sem=human, number=sing, deter=definite…]))
    NP[semrel=patient, logrel=arg1, synfunc=obj1, num=sing…]
      (adjposs[lex=‘his’, semrel=poss, synfunc=det, number=plur, deter=definite…]
       Noun[lex=‘money’, synfunc=head, number=sing…])
    NP[semrel=dest, logrel=arg2, synfunc=subj, sem=human, number=plur…]
      (Noun[lex=‘customer’, synfunc=head, sem=human, number=plur, deter=definite…]))
```

Abstract representations of utterances are far superior to concrete representations as input and output structures of transfers in semantic transfer MT or as “lexical-conceptual structures” [23] in interlingual MT, especially between distant languages. But their relation to the corresponding utterances is not as clear, a natural consequence of abstraction. That “remoteness” is even more apparent with other types of structures, such as conceptual graphs, logical formulae or interlingual representations *à la* KBMT-89 [27]. By contrast, concrete structures are clearly more adequate for interactive disambiguation. They are also superior for a variety of future applications. For example, no text processor today is able to replace "give-back" by "return" in the preceding example, not to speak of changing the modality, the tense, the voice or the discourse type (say, from direct to indirect or affirmative to negative). Self-explaining documents would make that possible.

Here is a functional diagram (figure 22 - 1) of the processes we have discussed above. In gma-structures (generating, multilevel and abstract), non-interlingual linguistic levels are underspecified and, if present, are used only as reflections of corresponding surface levels in the source language and are recomputed in the first generation phase, which we call "paraphrase choice".

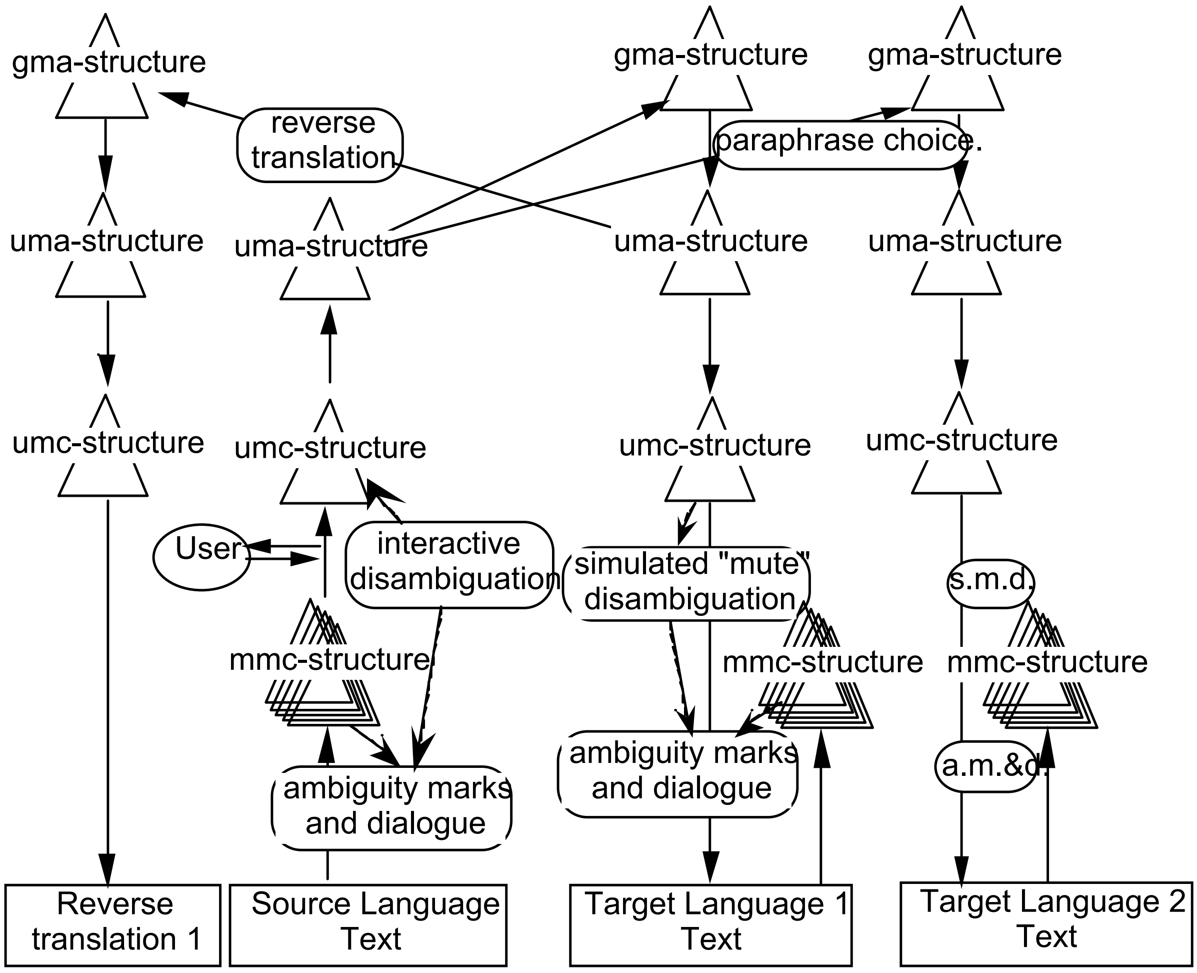

Figure 22 - 1

4 Alternatives

Is DBMT really a better approach than other alternatives? Our answer is a definite yes, because:

- very often, these alternatives are not really feasible,
- the results can be intrinsically better, if presented as self-explaining documents.

First, *machine aids for translators* [1, 21, 25, 26, 34] are usable only if there are available and affordable translators and if they have enough time to do the job. But, in many situations, there are simply no such translators, especially if translations are required in several languages. For example, multinational firms, banks, etc., have many uncovered translation needs. In Europe, scientists and engineers are engaged in many projects where communication is hampered by the language barrier.

Even if competent translators are available, the delays are often such that translation is impossible. That is for example the case in European institutions, which are theoretically required to issue all important documents in all official languages but are unable to do so, although they employ more than 1200 full-time translators and translate more than 1.2M pages a year. But the final versions of these documents are too often ready at the last moment. Here, it would make more sense to analyze and disambiguate their parts as soon as they are ready and to translate them at the last moment.

Another possibility, often advocated, is to write in *controlled languages* designed to be unambiguous and use “black box” MT. This can be very successful in restricted situations as in the case of the TITUS system [14] of the Institut Textile de France. But it is very difficult to force people to write in a controlled language. It proved for instance impossible to adapt the TITUS system to the context of the CDST (Center for Scientific and Technical Documentation of CNRS). Another weak point of controlled languages is that they are difficult to design and very task- and domain-specific.

Finally, controlled languages are unambiguous for the analyzer designed to process them, but not for humans. While this may be convenient in the context of man-machine communication, it may be counterproductive, or even dangerous, in the context of human communication. If, for example, “to replace (a mechanical part)” is intended to mean only “to replace by a new thing” (“remplacer”), and not “to put back in place” (“replacer”), a mechanic may well understand the second, unintended meaning, and put back in place an airplane part which should be replaced by a new one, leading to an accident.

If the concept of self-explaining document may be made to work in broad contexts, texts could be written without restrictions stronger than the usual ones which concern style and terminology and at the same time be in effect less ambiguous than texts written in controlled languages. Even if translation is not an issue, then the availability of *“text explainers”* might be a major advance in document processing technology.

5 Perspectives

Full-scale, general DBMT systems “for everybody” would require extremely large grammatical and lexical knowledge bases. To cover a whole language, a lexical data base should contain of the order of 3 million terms, corresponding to 4 to 5 million monolingual acceptions, and perhaps to twice as many interlingual acceptions for systems designed to handle 10 to 20 languages.

The development costs are staggering and probably out of reach if conventional lingware engineering techniques are used. For instance, it has cost EDR (Tokyo) about 1200 man-years to develop 300K terms in 2 languages (200K terminological and 100K general), with the associated

640K interlingual concepts (200K terminological and 2*300K general, minus 60K common). At least 100 times more (1.2M man-years!) would be needed for 3M terms in 20 languages.

This is why we advocate a step by step approach, and the development of new groupware techniques for developing very large lexical data bases.

First, text explainers could be developed for several languages, in specific domains and situations. Here, the figures would be 10 to 30K terms, 100 to 300 times less than those mentioned above.

Second, the (monolingual) analyzers and dictionaries developed for text explainers might be reused for building DBMT systems for the same restricted domains and situations. Transfers and generators would not be too costly to develop. The lexical work would consist in integrating all monolingual lexical data bases in a unique multilingual data-base, thereby refining each monolingual data-base according to the obtained set of interlingual acceptions [30].

In a third step, general text explainers and DBMT systems could then be developed, by progressively merging and extending specific systems. This "divide-and-conquer" strategy might break down the cost and make the whole enterprise feasible in the long run.

Even if the cost were not a problem, developing extremely large lexical data bases and keeping them up to date would be impossible using only professional teams of lexicographers. The quantity is too huge, changes and innovations are too fast, and only specialists can be competent in specific domains. We consider it as a major challenge to develop *groupware lexical data base development techniques* which might be based on the *contribution of lexical information by users* of existing text explainers or DBMT systems.

A distributed architecture would be very advantageous for that purpose, because lexical information created by the users on their personal computers might transparently and automatically be sent to the servers. For example, authors might add new senses to existing terms, add new terms, and propose translations in the languages they know. Professional teams would then process and refine this "raw lexical material" on server sites. This idea is not so far-fetched, and very similar to that used in Eurolang Optimizer™ [1], a distributed environment designed for professional translators.

References

[1] **Bachut D. (1994)** *Le projet EUROLANG : une nouvelle perspective pour les outils d'aide à la traduction.* Proc. TALN-94, journées du PRC-CHM, Marseille, 7—8 avril 1994, Univ. de Marseille.

[2] **Blanc É., Sérasset G. & Tchéou F. (1994)** *Designing an Acception-Based Multilingual Lexical Data Base under HyperCard: PARAX.* Research Report, GETA, IMAG (UJF & CNRS), Aug. 1994.

[3] **Blanchon H. (1992)** *A Solution to the Problem of Interactive Disambiguation.* Proc. COLING-92, Nantes, 23-28 July 1992, C. Boitet, ed., vol. 4/4, pp. 1233-1238.

[4] **Blanchon H. (1994)** *Perspectives of DBMT for monolingual authors on the basis of LIDIA-1, an implemented mockup.* Proc. 15th International Conference on Computational Linguistics, COLING-94, Kyoto, Japan, 5-9 Aug. 1994, vol. 1/2, pp. 115—119.

[5] **Boitet C. (1989)** *Motivation and Architecture of the LIDIA Project.* Proc. MTS-II (MT Summit), Munich, 16-18 août 1989, pp. 50—54.

[6] **Boitet C. (1990)** *Towards Personal MT : on some aspects of the LIDIA project.* Proc. COLING-90, Helsinki, 20-25 août 1990, H. Karlgren, ed., ACL, vol. 3/3, pp. 30-35.

[7] **Boitet C. (1993)** *La TAO comme technologie scientifique : le cas de la TA fondée sur le dialogue.* In "Études et Recherches en Traductique", A. Clas & P. Bouillon, ed., Presses de l'Université de Montréal, Montréal, pp. 109—148.

[8] **Boitet C. (1994)** *Dialogue-Based Machine Translation and Sub-Languages.* Proc. ICLA-94, Penang, Malaysia, 26-28 July 1994, USM.

[9] **Boitet C. & Blanchon H. (1993)** *Dialogue-Based MT for Monolingual Authors and the LIDIA project.* Proc. NLPRS'93 (Natural Language Processing Rim Symposium, Fukuoka, 6-7/12/93, H. Nomura, ed., Kyushu Institute of Technology, pp. 208—222.

[10] **Boitet C. & Blanchon H. (1994)** *Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup.* Machine Translation. (to appear).

[11] **Brown R. D. (1989)** *Augmentation.* Machine Translation, 4, pp. 1299-1347.

[12] **Brown R. D. & Nirenburg S. (1990)** *Human-Computer Interaction for Semantic Disambiguation.* Proc. COLING-90, Helsinki, 20-25 août 1990, H. Karlgren, ed., ACL, vol. 3/3, pp. 42-47.

[13] **Chandler B., Holden N., Horsfall H., Pollard E. & McGee Wood M. (1987)** *N-tran Final Report.* Alvey Project, 87/9, CCL/UMIST, Manchester.

[14] **Ducrot J.-M. (1988)** *Le système TITUS IV.* In "Traduction Assistée par Ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires", A. Abbou, ed., Observatoire des Industries de la Langue (OFIL), Paris, mars 1988, pp. 55—71.

[15] **Huang X. M. (1990)** *A Machine Translation System for the Target Language Inexpert.* Proc. COLING-90, Helsinki, 20-25 Aug. 1990, H. Karlgren, ed., ACL, vol. 3/3, pp. 364-367.

[16] **Hutchins W. J. (1986)** *Machine Translation : Past, Present, Future.* Ellis Horwood, John Wiley & Sons, Chichester, England, 382 p.

[17] **Kay M. (1973)** *The MIND system.* In "Courant Computer Science Symposium 8: Natural Language Processing", R. Rustin, ed., Algorithmics Press, Inc., New York, pp. 155-188.

[18] **Kay M. (1980)** *The Proper Place of Men and Machines in Language Translation.* Research Report, CSL-80-11, Xerox, Palo Alto Research Center, Oct. 1980.

[19] **Kittredge R. (1983)** *Sublanguage — Specific Computer Aids to Translation — a survey of the most promising application areas.* Contract n° 2-5273, Université de Montréal et Bureau des Traductions, mars 1983, 95 p.

[20] **Kittredge R. (1986)** *Analyzing Language in Restricted Domains.* In "Sublanguage Description and Processing", R. Grishman & R. Kittredge, ed., Lawrence Erlbaum, Hillsdale, New-Jersey.

[21] **Langé J.-M. (1994)** *Systèmes d'aide à la traduction : un point de vue industriel.* Proc. TALN-94, journées du PRC-CHM, Marseille, 7—8 avril 1994, Univ. de Marseille.

[22] **Lehrberger J. & Bourbeau L. (1988)** *Machine Translation. Linguistic characteristics of MT systems and general methodology of evaluation.* John Benjamins, 240 p.

[23] **Levin L. & Nirenburg S. (1994)** *The Correct Place of Lexical Semantics in Interlingual MT.* Proc. 15th International Conference on Computational Linguistics, COLING-94, Kyoto, Japan, 5-9 Aug. 1994, vol. 1/2, pp. 349—355.

[24] **Maruyama H., Watanabe H. & Ogino S. (1990)** *An Interactive Japanese Parser for Machine Translation.* Proc. COLING-90, Helsinki, 20-25 août 1990, H. Karlgren, ed., ACL, vol. 2/3, pp. 257-262.

[25] **Melby A. K. (1981)** *Translators and Machines - Can they cooperate ?* META, **26**/1, pp. 23-34.

[26] **Melby A. K. (1982)** *Multi-Level Translation Aids in a Distributed System.* Proc. COLING-82, Prague, 5-10 juillet 1982, vol. 1/2, pp. 215-220.

[27] **Nirenburg S. (1989)** *Knowledge-based Machine Translation.* Machine Translation, 4, pp. 5-24.

[28] **Nyberg E. H. & Mitamura T. (1992)** *The KANT system: Fast, Accurate, High-Quality Translation in Practical Domains.* Proc. COLING-92, Nantes, 23-28 July 92, C. Boitet, ed., ACL, vol. 3/4, pp. 1069—1073.

[29] **Sadler V. (1989)** *Working with analogical semantics : Disambiguation technics in DLT.* T. Witkam, ed., Distributed Language Translation (BSO/Research), Floris Publications, Dordrecht, Holland, 256 p.

[30] **Sérasset G. (1994)** *Interlingual Lexical Organisation for Multilingual Lexical Databases.* Proc. 15th International Conference on Computational Linguistics, COLING-94, Kyoto, Japan, 5-9 Aug. 1994, 6 p.

[31] **Sérasset G. (1994)** *An Interlingual Lexical Organization Based on Acceptions.* Proc. ICLA-94, Penang, Malaysia, 26-28 July 1994, USM, 12 p.

[32] **Somers H. L., Tsujii J.-I. & Jones D. (1990)** *Machine Translation without a source text.* Proc. COLING-90, Helsinki, 20-25 Aug. 1990, H. Karlgren, ed., ACL, vol. 3/3, pp. 271-276.

[33] **Tomita M. (1986)** *Sentence Disambiguation by asking.* Computers and Translation, **1**/1, pp. 39-51.

[34] **Tong L. C. (1987)** *The Engineering of a Translator Workstation.* Computers and Translation, **2**/4, pp. 263—273.

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

[36] **Wehrli E. (1992)** *The IPS System.* Proc. COLING-92, Nantes, 23-28 July 1992, C. Boitet, ed., vol. 3/4, pp. 870-874.

[37] **Whitelock P. J., Wood M. M., Chandler B. J., Holden N. & Horsfall H. J. (1986)** *Strategies for Interactive Machine Translation : the experience and implications of the UMIST Japanese project.* Proc. COLING-86, Bonn, 25-29 août 1986, IKS, pp. 25-29.

[38] **Wood M. M. G. & Chandler B. (1988)** *Machine Translation For Monolinguals.* Proc. COLING-88, Budapest, 22-27 Aug. 1988, pp. 760—763.