Evolution of MT with the Web

Christian Boitet[1], Hervé Blanchon[1], Mark Seligman[2], Valérie Bellynck[1]

{Christian.Boitet, Herve.Blanchon, Valerie.Bellynck}@imag.fr, Mark.Seligman@spokentranslation.com

[1]GETALP, LIG-campus
(UJF, CNRS, INPG, INRIA, UPMF)
BP 53, 38041 GRENOBLE cedex 9, France

[2]Spoken Translation, Inc.
1100 West View Drive
Berkeley, CA 94705

International Conference "Machine Translation 25 Years On", Cranfield, England, 21-22/11/2009

Abstract

Since the Cranfield-94 conference, we have come to a better understanding of the nature of MT systems by separately analyzing their *linguistic*, *computational*, and *operational architectures*. Also, thanks to the *CxAxQ metatheorem,* the systems' inherent limits have been clarified, and design choices can now be made in an informed manner according to the translation situations. *MT evaluation* has also matured: tools based on reference translations are useful for measuring progress; those based on subjective judgments for estimating future usage quality; and task-related objective measures (such as post-editing distance) for measuring operational quality. Moreover, the same technological advances that have led to "Web 2.0[1]" have brought several futuristic predictions to fruition. Free Web MT services have *democratized assimilation MT* beyond belief. *Speech translation* research has given rise to usable systems for restricted tasks running on PDAs or on mobile phones connected to servers. New man-machine interface techniques have made *interactive disambiguation* usable in large-coverage multimodal MT. Increases in computing power have made *statistical methods* workable, and have led to the possibility of building low-linguistic-quality but still useful MT systems by machine learning from aligned bilingual corpora (SMT, EBMT). In parallel, progress has been made in developing *interlingua-based MT systems,* using *hybrid methods.* Unfortunately, many *misconceptions about MT* have spread among the public, and even among MT researchers, because of ignorance of the past and present of MT R&D. A compensating factor is the *willingness of end users to freely contribute* to building essential parts of the linguistic knowledge needed to construct MT systems, whether *corpus-related* or *lexical.* Finally, some developments we anticipated fifteen years ago have not yet materialized, such as online writing tools equipped with interactive disambiguation, and as a corollary the possibility of transforming source documents into *self-explaining documents* (SEDs) and of producing corresponding SEDs fully automatically in several target languages. These visions should now be realized, thanks to the evolution of Web programming and multilingual NLP techniques, leading towards a true Semantic Web, "Web 3.0," which will support "ubilingual" (ubiquitous multilingual) computing.

Keywords

MT, linguistic architecture, computational architecture, operational architecture, task-related evaluation, speech MT, interactive disambiguation, self-explaining documents, Semantic Web MT

Introduction

This paper addresses the call to assess the achievements of Machine Translation (MT) and Machine Assisted Translation (MAT) in the twenty-five years since 1984 and to look ahead toward expectations for R&D in MT and MAT in the next twenty-five years. More than four authors would certainly be required to present the last twenty-five years of MT completely and faithfully, and doubtless no combination of experts could actually predict what will happen by 2034. We will nevertheless risk some predictions, knowing full well that only about half of our predictions in previous papers [30, 32, 36] have come true. We also know that there have been some altogether unexpected developments, such as the growth of empirical methods (SMT, EBMT). As in the field of speech recognition, these methods became practical because of continuing increases in computing power and storage capacity. In addition, the Internet and the free, open-source spirit of the Web have given access to very large linguistic resources: users can now help to build them and can improve MT results online, using free Web browsers. In these efforts, users now experience no more delay than when using traditional and often expensive MT and TA (Translation Aid) systems on a PC.

---

[1] For definitions, see http://www.atelier-informatique.org/internet/evolution-web-10-web-20-web-30/358/. "Web 1.0" refers to the time when all pages were static. "Web 2.0" techniques allow for data sharing, dynamic pages, social networks, and collaborative work, and include the exploitation of XML, RSS and AJAX. Users become active participants. Web 3.0 has just begun (from 2008), the main differences being knowledge processing *(Semantic Web, intelligent Web),* P2P *(peer to peer)* communications and services, and personalization (facial recognition, net identity, etc.). As computing becomes pervasive*,* multilingual computing should become *ubilingual (ubiquitous multilingual) computing.*

One particular development has been the incredibly rapid progress of speech translation. The first three authors have been involved in this work, and one (Seligman) has in fact built a commercial, large-coverage PC-based system, Converser for Healthcare™, fulfilling a vision which we three developed at ATR in 1992-93: we anticipated that interactive disambiguation, added to a combination of commercially available ASR, MT and TTS[2] modules, could lead to the creation of practical systems. The fourth author adds a view of the potential of Web 3.0 techniques, having participated in the design of user-centered microworld interfaces; of online tools eliciting lexical contribution in a learning context; of modern techniques of Web programming; and of interactive multilingual translation gateways.

In the first section of this paper, rather than recount all of the turns of MT since 1984, we will summarize the progress we observe in understanding MT, TA, and their interrelations since that year. In Section 2, we cover aspects of the evolution of the MT field that seem related to the evolution of the Internet. We start with the last Cranfield conference (1994), because the prior period has largely been covered elsewhere: see in particular [34, 71, 72]. That limitation provides a reason to limit our lookahead in Section 3, where we describe the MT evolution we expect in the next fifteen (rather than twenty-five) years, in the context of the Semantic Web (Web 3.0).

1. Progress in understanding MT, TA, and their interrelations since 1984

1.1 Dimensions of analysis of MT systems architectures

MT systems were first classified according to the technological approach used: in 1976, one spoke of "nG" (n-th generation) systems. 1G systems (e.g., GAT, Systran) were *direct* (word-for-word) and *directly* programmed; 2G systems used *intermediate linguistic structures* and were programmed in *metalanguages*; and futuristic 3G systems would translate by *understanding,* using world (*gnostic* or *encyclopedic)* knowledge, and would to some degree mimic translators' brain processes.

After 1980, MT systems began to be classified according to their immediate users: *MT for watchers, MT for revisors, MT for translators,* and (around 1989) *MT for authors*. Since 2000 or so, one reads about the opposition between *rule-based MT* (RBMT) and *statistical MT* (SMT), though these terms really distinguish between *handcrafted MT* (HCMT) and *Machine-Learned MT* (MLMT).

Since the Cranfield conference in 1994 ("MT Ten Years On," meaning ten years after the first conference in 1984), the first author has tried to propose a more exact classification. First, a distinction was made [50] between two aspects of the analysis of MT systems: their *linguistic architecture* and their *computational architecture.* The need to consider a third aspect, their *operational architecture*, was then recognized.

The linguistic architecture may be defined by the succession of intermediate representations (IR) used to process a translation unit (the path in Vauquois' triangle, see Figure 2), as well as their precise nature and scope (only sentences, or paragraphs, or even full texts).

The computational architecture is defined by the technology used to build the *phases* transforming one IR into the next. A phase may be hand-crafted (using classical programming languages, or rule-based languages, of various types[3]), or it may be more or less automatically learned from bilingual examples (pairs of strings, of trees, of <string, tree> pairs, or of <string, abstract representation> pairs).

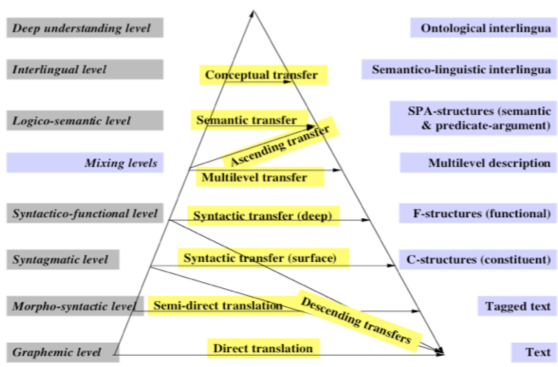

Figure 1: linguistic architectures of MT systems

The terms *expert* and *empirical* have been proposed recently for expressing this hand-crafted vs. machine-learned distinction, and we will use them here.

[2] ASR = Automatic Speech Recognition, TTS = Text To Speech (for voice synthesis).

[3] Rules may be non-procedural rules of well-formedness (as in formal grammars like CFGs, TAGs, or HPSGs), or they may be procedural rules (rewriting rules on strings, trees or graphs, and transitions of automata such as FSTs).

A detailed study of current and past MT systems shows that *the linguistic and computational architectures of MT systems are independent* of each other (Boitet 2007).

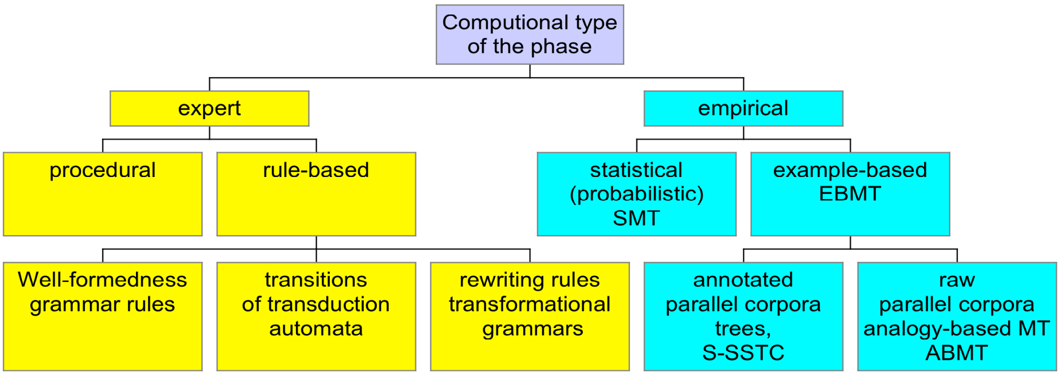

Figure 2: hierarchy of computational architectures of phases of MT systems

By *operational architecture,* we mean the precise conditions of the use and development of a system:

- *Tasks and users:* (1) help bilinguals produce good translations; (2) help people understand an unknown or little-known language; (3) help people communicate (chat, spoken translation…).
- *Language pairs / volumes / kinds:* graph of translation paths or directions (1→1 as in ALT/JE; 1↔1 as in Converser for Healthcare; 1→N as in MedSLT; 1↔N as in Phraselator (for the US Army); or N↔N (for debates, multilingual chat, multilingual peace forces)
- *Possible involvement of humans:* (1) authors (controlled language, rewriting, interactive disambiguation); (2) professional/occasional translators (post-editing); and (3) readers (guessing from multi-result factorized output).
- *Available resources:* data (in particular, the huge parallel corpora necessary for SMT), and humans (computational linguists, lexicographers needed for "expert" MT).

1.2 The CxAxQ MT (meta)theorem

Theorem Statement: *The product of language Coverage, Automation rate, and linguistic Quality of MT systems is always well below 100%, but two of these factors can approach 100% if one compromises on the third.* As a formula, **Coverage x Automaticity x Quality << 100%.** That limitation is in the nature of the problem.

This statement is experimentally but not formally provable, exactly like Church's Thesis. Thanks to it, the inherent limits of translation automation have been clarified, and design choices can be made in an informed manner according to the translation situation. For example, when large coverage and high quality are needed, as for helping English-speaking health personnel to converse about virtually anything with Spanish-speaking patients and their families, automation must be far less than 100%: we have to let users disambiguate and give them considerable control – as in Converser for Healthcare, which supports reverse translation and optional user-initiated lexical disambiguation. Similarly, if we aim at very high quality and full automation, we can only build MT systems tailored to a restricted sublanguage, such as weather bulletins [60] or stock market flash reports [90]. In the case of Web translations, C x A ≈100%, so quality simply cannot be 100%; but C x Q ≈100% in the case of DBMT systems such as KANT/CATALYST [84] or LIDIA [9]. *Quality* usually means linguistic quality, as judged by translators. We propose to define it relative to the post-edition time, expressed in minutes per page, so that $Q \leq 0$ if more than 50 minutes[4] are needed to post-edit the result of one standard source page (250 words): Q=(100-2xT(post-edition_MT))%.

Post-edition T (minute/page)	0mn	5mn	10mn	15mn	20mn	25mn	30mn	35mn	40mn	45mn	50mn	55mn	60mn	65mn
MT quality (%)	100%	90%	80%	70%	60%	50%	40%	30%	20%	10%	0%	-10%	-20%	-30%

[4] (Linguistic) quality will be 0% if the MT result is so bad that post-editors reject it and restart from scratch, even if they can gain ten min/page by using it. Linguistic quality is different from usage quality (i. e, utility, or cost-effectiveness).

To define *automaticity*, we propose the formula $A=1-\frac{\text{Time(human\_interaction)}}{\text{Time(human\_first\_draft)}}$ which takes into account all kinds of human intervention in the translation process, rather than only post-edition time. For example, A = 83.3% if one needs only ten minutes of human interaction with an MT system to produce the draft translation of a page, normally produced in one hour of human work.

In [51], we give a method of computing the *coverage* of an MT system, but cannot present it here for lack of space[5].

1.3 What to evaluate in MT and how to do it

MT evaluation methods can be *internal* (used while viewing a system's inner mechanisms in order to develop or technically assess the system) or *external* (viewing the system as a black box). Internal measures have received much less attention from researchers than external measures, probably because they must be at least partly system-specific. Nevertheless, all MT developers and editors have their own test suites of sentences and documents.

With respect to MT evaluation, the most important development since 1994 is the introduction in 2001-2002 of *objective external measures* such as BLEU, NIST, ORANGE, WER, mWER, etc., based on automatic comparisons of MT results with *reference translations*. (Most such measures use only easily computable n-grams counts.) Competitive MT evaluation campaigns have been organized based upon these measures, comparable to those for speech recognition, information retrieval, Q&A, etc. These campaigns have certainly been worthwhile. However, the expectation that these n-gram-based measures would correlate well with human judgments of linguistic quality has not been met [59]. Moreover, to evaluate systems currently in use, reference-based measures require human intervention: by definition, there are no reference translations of previously untranslated segments. In order to use these measures, then, one should modify the operational context and architecture so that humans can post-edit MT results, producing reference translations of a fraction of the input, as in Google Translate. But even in such a favorable case[6], the measure is certainly biased, because the sample of segments selected by web surfers for contributive post-edition is not at all random.

Our understanding of MT evaluation techniques has now matured. In [12, 15], we conclude that it is best to evaluate operational quality by using *task-related measures* which compare (a) the effort to perform a certain task with the MT system, integrated in its operational environment, and (b) the effort without that system. One should limit the use of tools like BLEU to measure progress during MT system development. However, contrary to an all-too-common belief, these tools are not inherently limited to empirical (auto-learned) systems (SMT and EBMT). It is true that, in the case of expert (handmade) MT systems, reference translations given from outside are mostly inadequate, because they are too far from actual MT results (in the set of all possible target sentences): in general, much nearer reference translations may be found, given the immense number of acceptable translations of a segment. Thus, for expert MT systems, the solution is to post-edit the MT results on test suites and then to use the corrections as reference translations. In summary, (1) tools based on reference translations are useful for measuring progress; (2) those based on subjective judgments are useful for estimating future usage quality; and (3) task-related objective measures (such as post-editing distance) are useful for measuring operational quality.

1.4 Good surprises

Web translation. The same technological advances that have led to Web 2.0 have brought several futuristic predictions to fruition. Free Web MT services have *democratized assimilation MT* beyond belief.

Speech translation. Research has led to creation of fully automatic but usable systems for restricted tasks, running on PDAs or on mobile phones connected to servers. Back in 1993, many thought this effort could never succeed because ASR and MT errors would multiply. But tremendous progress has been made on the ASR side – recognition is now possible under noisy conditions without sophisticated microphones – and new

---

[5] Briefly, we prepare a table with rows for the types of linguistic phenomena (words/compounds, terms, collocations, grammatical constructions), and columns for frequency and importance (FI) relative to the translation situation, and coverage of general meanings (GM) and specific meanings (SM). Hence, coverage should depend on the identified task.

[6] That has become possible only since the introduction of the contributive Web (Web 2.0).

man-machine interface techniques (on PCs, PDAs, or even mobile phones) allow user control and participation, in particular by correction of ASR outputs and interactive disambiguation in the source language [87, 88]. Broad-coverage multimodal MT is emerging.

Empirical and notably statistical methods. Increases in computing power have made *statistical methods* workable, and have made possible the construction of low-linguistic-quality but still useful MT systems by machine learning from aligned bilingual corpora (SMT, EBMT).

Hybridization. The MSR-MT system developed by Microsoft Research around 1998-2002 [63] used hand-crafted analysis and generation, but learned the transfer component totally automatically by compiling a MindNet from a large number of corresponding English and Spanish trees. Combination of empirical methods with interlingua-based linguistic architectures has been demonstrated, notably by the IBM MASTOR hand-held English-Mandarin speech translation system [66, 94]. All of this work demonstrates the possibility of developing *interlingua-based MT systems* using *hybrid methods.* Another aspect of hybridization is the use of corpus-crunching methods to prepare data for expert (handmade) MT systems. For example, the size of the Fujitsu ATLAS-II dictionaries has grown from 586,000 entries in 2001 (at that year's MT Summit) to more than 5.57M in 2008 (v.14). No doubt this increase was achieved not only through manual creation or adaptation of existing dictionaries, but also by exploiting parallel or comparable corpus processing techniques. Very large increases in dictionary size can also be observed in The Honyaku (Toshiba), Systran, etc. Likewise, most TA tools (Trados, XMS, Similis…) now include good or excellent terminology extractors.

1.5 Increase of misconceptions about MT

MT has always been a somewhat mythical field, with many misconceptions, apocryphal examples, etc. However, this tendency has increased rather than lessened in the past twenty-five years. In particular, many MT researchers are ignorant about the history of MT, the nature of MT, and the existing operational MT systems. Here are examples of such misconceptions.

The majority of operational systems use the SMT design. No. Google Translate and LanguageWeaver (which offers no free Web server) are indeed SMT-based systems, but Systran, LMT, METAL, ProMT, Reverso, WordMagic, ATLAS, The Honyaku, ALT/JE, ALTFLASH, Neon, etc. are totally or essentially hand-crafted. However, it is true that some of these latter systems (like Systran) are beginning to insert a statistical (or rather, probabilistic) phase into their computational architecture.

Pivot implies Rule-Based. No, as demonstrated by the MASTOR system (2003), and previously by a part of the multilingual, CSTAR-II spoken dialogue MT system (1999) based upon the IF pivot standard.

Pivot implies Interlingua. No: a pivot can be a structure based upon some representational level of a particular natural language, e.g. analysis results in multitarget MT. That is the case in the MedSLT systems, and also in the LIDIA prototype mentioned above.

MT with interlingual (IL) pivot cannot work and scale up. No. ATLAS-II (Fujitsu) has been the best system for Japanese to and from English for twenty years. ATLAS has more than 5.7M dictionary entries (v.14, Dec. 2008).

Transfer MT with N languages implies N(N-1) transfers. No. One can use the analysis representations of some language as intermediate structures and can then combine two transfers to obtain translations for (N-1)(N-2) language pairs. Only 2(N-1) transfer components are needed in this case, a linear rather than quadratic number.

Statistical MT implies low development cost. Not if one takes into consideration the cost of producing the bilingual corpora. For example, one parallel corpus of 50M words (200K pages) has required between 150K and 200K hours of human translators (about 100 person-years) for its production. While it is true that an SMT system can be produced virtually overnight once suitable corpora are in place, LanguageWeaver produced only four systems between 2001 and 2005, most probably because of the scarcity of clients having large enough previously translated corpora. It should also be noted that large corpora cannot easily be borrowed: an SMT system developed for one kind of corpus will produce far worse results on another kind of corpus.

Rule-Based MT implies high development cost. No – or rather, it depends. True, specialists are needed to design and implement the expert lingware components (grammars, automata, and dictionaries), and their work does cost more than outsourced manual translation. However, the development of systems aiming at sublanguage, such as METEO or ALTFLASH, has been very cheap (3-5 persons for 6-10 months). For large-coverage systems, Fujitsu reported a cost of 300 man-years to market the first versions of ATLAS-II (around 1982), with 70,000 entries in each dictionary. By way of comparison, if a parallel corpus of 200 M words is needed to obtain comparable output quality, as suggested by P. Koehn at a meeting of the UE in 2006, the cost would be from 600 to 800 man-years.

MT quality has increased with SMT. No. On the contrary, the very design of SMT systems lowers the asymptotic (best possible) quality. In 2004-05, the RALI (Université de Montréal) built an SMT system from an aligned corpus of 40 M words of weather bulletins translated by METEO and revised by experts [76]. They obtained 77% acceptable translations, so at least 23% corrections would have to be made, while only 3% were actually made on the METEO results.

BLEU measures the quality of translations. No (see above). Many researchers are aware that BLEU scores do not correlate well with human judgments of translation quality, but still use this tool as a quality measure in research papers because it is available and simple to use. That is a major flaw in current MT evaluations.

Adequacy should be measured by a positive number. No. In fact, information conveyed by a translation can be actually misleading: a "contresens" (mistranslation) can lead to false assumptions. We have seen many such cases while participating in the subjective evaluation of IWSLT campaigns. Thus adequacy should rather be computed using kappa coefficients, which measure agreement or disagreement between the original message and a candidate translation.

1.6 Emergence of theoretical arguments in favour of example-based approaches

The empirical approach – the idea of working with examples rather than full-blown theories – has been defended mainly through practical rather than theoretical arguments. In the past, a similar situation obtained in the area of speech recognition: back in 1975, CMU had developed a knowledge- and theory-based system (Hearsay), and in parallel the empirical HHM-based Harpy system. Harpy won, by a large margin. Later, Jelinek famously said that whenever he fired a linguist his ASR recognition rate would jump up 1%. The same argument has more recently been used in relation to SMT: it works, so what's the problem? The general problem is that SMT does not in fact work better that expertly built systems. Among other more specific problems, new kinds of very annoying errors appear: parts of the input message are left out; parts appear that don't correspond to anything in the source; and a study by Dong Zheng Dong shows that overall reliability is considerably lower. Accordingly, we remain dissatisfied with the argument that decent results can justify the SMT approach without further ado. At the other extreme, we would be equally dissatisfied with the claim that SMT cannot work because it is not based on explicit deep knowledge. The fact is that empirical methods can perform to a surprising degree, but sometimes fail in the ways just mentioned and others; so theoretically oriented researchers should try to explain both the successes and failures.

Ed Hovy [69] has argued quite generally that, rather than develop a maximally complete and consistent theory, it would be more efficient to accumulate an enormous number of *factoids*, and to produce translations from them in MT (or find answers from them in information retrieval). We can only agree, as these areas are in effect *scientific technologies* [28] and not sciences. However, we think it is possible to go one step further.

Incompleteness and inconsistency metatheorem. Our claim: *there can be no satisfactory formal theory (axiomatization by axioms and rules) for an NL* (though there can be such a formal theory for a restricted enough controlled language, or for a restricted sublanguage which arises naturally). *Any axiomatization will undergenerate* (exhibit incompleteness) and *overgenerate* (exhibit inconsistency). Moreover, there seem to be standard ways to derive counterexamples from each proposed axiomatization. This is a game linguists like to play, and it justifies the never-ending quest for an adequate formal theory of NL. Our proposal, then, is that such a *formal* theory cannot exist, but a valid *semantic* theory can – as a set of valid sentences in a given language. That metatheorem is an analog of Gödel's incompleteness theorem: while there can be no axiomatic system (a *formal theory,* in mathematics) that generates all and only true statements about integers, the set of true statements about integers certainly exists (it is called the *semantic theory* of integers).

As a consequence, we are justified in abandoning the Holy Grail of building a perfect grammar and lexicon for a language, and in concentrating instead on known instances of the indescribable semantic theory – that is, on examples. That is exactly what happened with Vauquois & Chappuy's formalism for static grammars [91]: the formalism was used as a semi-formal specification level, up to the point where it became possible to automatically produce draft analyzers and generators using it [93]. From that point, however, it proved more efficient to build tools for empirically developing instances of the string-tree correspondences described by that type of grammar, the SSTCs [92], to align or *synchronize* them (obtaining a collection of synchronized SSTCs, or S-SSTCs), and to build an EBMT system from the aligned results [1]. Similarly, USM's EBMT system Banterjah has been obtained empirically, but its knowledge-base was produced in a hybrid way. The quality is said to be superior of that of an SMT built with the same size of initial bilingual corpus, but Banterjah uses a database of about 10K S-SSTCs, corresponding to about 600 pages of text in one language, semi-automatically prepared at a rate of about 1/2 hour per S-SSTC.

2. MT evolution since 1994, linked to the Web (1.0 and 2.0)

This section will be short, as there are not many disagreements here.

2.1 Democratization of access to Web translators

In 1994 (our starting date here), Systran was the first MT system to be made available on the Web, after having been on France's Minitel for ten years or more. Twenty-five years later, almost all MT vendors offer free Web page translators, and provide MT gateways like Voilà, Babelfhish, etc. to large customers. Millions of Web pages are translated on the fly every day, in more than fifty language pairs. This progress would of course have been unthinkable before the Web.

2.2 Building heterogeneous MT systems by collaborative development through the Web

The examples of UNL and CSTAR-II show that collaborative development via the Web is indeed possible. Intermediate structures encoded in XML or in HTML-like specific formats (as in `.unl` files, which contain interlingual representations in the format originated by H. Uchida [89]) are exchanged almost instantly between modules of MT systems running in various locations. These various structures are developed with very different tools and theoretical backgrounds. The availability and diffusion of XML, Unicode, and associated tools since 1998 has solved many difficult problems linked with the computer representation of texts in various languages.

2.3 Emergence of the application to MT of collaborative resource building

With the advent of Web 2.0 – the participative Web – the age of collaborative construction of resources in general, and of MT resources more specifically, is upon us. The idea of exchanging translation dictionaries, in particular, was pioneered by A. Melby before 1984, as soon as microcomputers became available [78, 79], with later developments towards the exchange of lexical and terminological data (via Micro-MATER [80]). More recently, collaborative human translation has become very active, with Web sites dedicated to the translation of documents related to causes (PaxHumana), or centers of interest (Wikipedia), or open source software (the W3C and Mozilla localization projects). Existing parallel HQ corpora like EuroParl and JR-acquis have become available on the Web in GPL[7], and more corpora are being created by volunteer translation communities. Cooperative development has begun more recently on (pre-terminological) dictionaries, especially in instructional contexts [5, 6].

2.4 Possibility of building reactive and contributive translation gateways

Due to Ajax techniques, which enable the programming of more dynamic Web pages providing richer and more participatory user experiences, it is now possible to browse a Web site in one's own language, modify a mistranslated segment on the fly, and continue reading. Behind the scenes, post-edited segments are stored in translation memories, and reused later if an exact match is found.

2.5 Possibility of quickly building MT systems for under-resourced languages or pairs

Even in the lucky case where a large enough parallel corpus is available for an under-resourced language pair (such as French-Thai or French-Vietnamese), these pairs pose special problems for empirical techniques, so that direct use of an SMT-building toolkit like Moses [85] is unsatisfactory. Language-specific preprocessing is needed, e.g. for segmentation in writing systems without word separators, for lemmatization for languages with complex inflectional morphology, and for word decomposition in languages with complex compositional morphology, as in Dravidian languages. As Martin Kay recently

[7] GNU Public License

observed, many of the wordforms found in the source text will never have been seen before. In addition, translation into languages with complex inflectional morphology often fails because of the lack of an underlying syntactic structure to guide generation. However, techniques have been designed to build these additional components by adapting components already available for other languages presenting similar problems. Such cloning techniques have been quite successful, for example, in producing ASR systems for Vietnamese, Khmer, Thai, etc. It also becomes possible to regenerate an MT system as time goes on, automatically adapting to changing elements like named entities and their translations. For example, a Vietnamese-French MT system for written news was built in 2009 from a Web site of translated news items [62]. A program first identifies translated pages using date proximity and proportion of recurring Vietnamese-French pairs (a dictionary or a phrase table may help) and thus produces a V-F corpus. It then extracts from the new corpus a parallel corpus of bi-segments, and finally generates an MT system using Moses. The quality of the results on similar untranslated news items seems to be better than that of Google Translate (which probably performs two-step translation through English).

3. MT evolution expected in Web 3.0

3.1 DBMT MT systems and Self-Explaining Documents (SED)

Back in 1994, we had just completed an experiment on a small DBMT (Dialogue-Based MT) prototype, LIDIA, translating from French into Russian, German and English. In the course of our experimentation, we observed once again that translation may introduce ambiguities which are not present in the source text. It also may happen that all of the disambiguated analyses of a sentence produce the same translation, which turns out to be 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*.

In that case (we may well ask), what is the use of disambiguating the source text if ambiguities only reappear in the translation(s), or – even worse – if new ones are created? Would it not be better to try to produce translations which preserve the ambiguities, thus dispensing with interactive disambiguation altogether? Unfortunately, experience with human translation shows that ambiguities can be *exactly* preserved only in some cases, and that to preserve them purposefully is quite difficult and often leads to unnatural expressions in the translation. 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, most ambiguities are not intentional, but are due to the intrinsic nature of natural languages. Some people write more clearly than others, but everybody writes ambiguously in all natural languages, which are ambiguous by nature (though unambiguously in all programming languages, which are unambiguous by construction).

Such considerations have led us to the idea of *self-explaining documents:* if the target documents are accompanied by their (unambiguous) linguistic structure, with 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 ambiguity issues are largely resolved. As human users are notably insensitive to ambiguities, however, we should find a way to warn the reader that the target text is ambiguous. In a multilingual DBMT setting, such warnings are possible. The system analyzes the target text with the analyzer of the target language, and produces a structure factorizing the multiple possible analyses, or *mmc-structure*[8]. It then automatically runs a simulated (and mute) disambiguation dialogue on the target side, automatically answering each disambiguation question so that the generated disambiguated structure, or *umc-structure*, is contained among the remaining candidate analyses as each question is answered. The system memorizes the disambiguation questions and the answers (Figure 3). It is then possible to show the ambiguities in the user interface by any convenient means, e.g. by creating buttons on which the reader may click to obtain the clarifications furnished by questions and answers – clarifications which *would have been given by the author, had the text been written in the target language*. In the Web 2.0 context, an SED is obtained by adding to an XML document and its folder of satellite files a *companion explanatory document,* also in XML. In [16, 61], we report on a prototype implementation of a *SED viewer and editor,* based on the Amaya XHTML editor of the W3C.

[8] mmc = multiple (all analyses), multilevel (from surface to abstract levels), concrete (direct correspondence with text). umc = unique (disambiguated), multilevel and concrete.

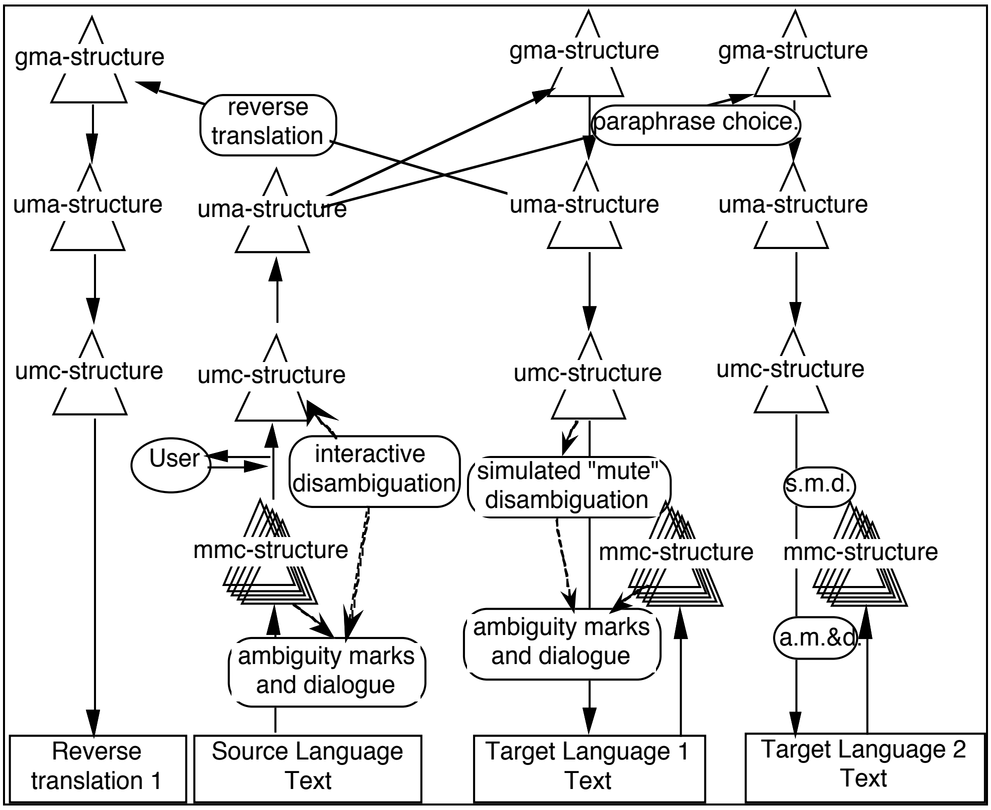

Figure 3: production of self-explaining documents, interactively in the source language, and then automatically in the target languages

Since 1994, we have been unable to obtain funding to develop a full prototype of this concept. Also turned down were even less ambitious project proposals to the EU concerning the introduction of writing tools linked to a DBMT system, although the potential of the approach for producing low cost, high quality translation into multiple languages had been amply demonstrated, in particular by the IBM JETS system. These concepts were simply too far ahead of their time.

With the advent of the Semantic Web, we hope they will no longer be seen as futuristic, but as quite feasible. It is now possible to develop such multilingual systems through the Internet. If a pivot-oriented approach is chosen, a team can be responsible for each language. Teams can share resources in a practical manner since, for example, sharing a common lexical database in real time no longer poses problems.

3.2 Interactive disambiguation in the target language

Some expert MT systems have been built to produce an ordered list of translations for each segment. The first translations are used to compose the output text, and the user interface allows an editor to see and select other translations further down the list. But post-editing the first version is almost always quicker than going down the list, often without finding any better preliminary translation. The situation is now far worse with SMT systems, which can produce very long scored lists (sometimes factorized in lattices of candidate translations): according to experiments performed at Xerox/XRCE, for 90% of all input sentences, there is no good translation in the first top 100 candidates.

The idea arises, then, of offering the editor a better structure to search through. Rather than a list, some sort of graph seems indicated, in the manner of a confusion network, offering readers a convenient interface to quickly search the network for a translation meeting their expectations. In addition to facilitating navigation through numerous candidate translations produced by a single MT system, such an approach could also enable the easy combination of parts of candidates from several MT systems. Note that, in this scenario (which might motivate a new workflow design, or operational architecture), users are expected to know the domain and the target language very well, and the source language minimally or not at all. The approach can be implemented via a standard Web browser, as demonstrated by [86]. In this implementation, the text of the translation's current version appears in the interface's central frame. It looks like normal text, and is composed of the currently selected trajectories in the factorizing graph. The text is clickable where there are alternations in the graph. The interface's left pane is used to manipulate the graph (e.g. to change a selection or edge priority, or to indicate crossings or additions). We think this new concept of interactive translation exploiting target-language graphs may become possible and useful in the Web 3.0.

3.3 Knowledge-Based MT

The idea of connecting an MT system to a knowledge base through an expert system was first studied and prototyped by R. Gerber in his 1984 PhD research [67]. The purpose was to solve certain lexical ambiguities and to correct errors in parse trees, in particular attachment errors. Later, immediately after MT research resumed in the US after eighteen years in disrepute, CMU [82] proposed to build an MT system around a full-blown domain ontology, ultimately producing the KBMT-89 prototype. The KANT system was then developed for Caterpillar and deployed there with four target languages. Still later, after 2000, it appears that a TA system, Déjà Vu, was put into operation at Caterpillar. We could not determine whether MT continues there, perhaps integrated with TA; however, it would not be surprising if it has been discontinued: it is often too costly to maintain MT dictionaries in synch with those for a translators' assistant, which have been enriched by translators. Highly specialized linguists must in this case follow the evolution of input sublanguage, and must develop and maintain more target languages. In the end, MT usage becomes less and less attractive and is finally discontinued.

The old concept of KBMT is now bound to revive, we believe. As *domain ontologies* are becoming popular, they will not have to be built for the sake of MT. (By contrast, ontology construction more than doubled the cost of the lexical knowledge in KBMT-89.) Rather, ontologies will be employed as they are, using new sorts of expert system interfaces, comparable to those R. Gerber used twenty-five years ago. For these purposes, it will be possible to use a new technique we are developing (in the context of the French ANR OMNIA project in order to access large databases of images with companion texts). Descriptors relative to a domain ontology Ω are extracted by image processing and textual content extraction, then merged and stored in the facts, or *A-box*, of ontology Ω. We *multilingualize* context extraction in two steps: first, we annotate the texts by interlingual lexemes (such as the UWs of UNL), using a language-specific module; next, we run a language-independent content extraction algorithm on the interlingual annotations, guided by an automatically built correspondence between the interlingua and the concepts, attributes, and rules, or *T-box*, of ontology Ω.

Conclusion

MT has changed considerably since 1984, and the pace has quickened since 1994. We will not repeat here the summary in the abstract above. Rather, we will conclude by saying that the most important factors for success or failure of MT systems in the future will probably be the adequacy of their operational architecture (workflow design, users, contributors and scenarios of use) to support the new uses which will appear with the Semantic Web, Web 3.0. In particular, the development of high quality MT for a very large number of naturally arising sublanguages will necessitate the extensive involvement of users, both as volunteer co-developers of resources and as improvers (post-editors) of rough or raw MT results. Widespread high-quality MT development will also require good methods for extending text-based and spoken MT to under-resourced languages and language pairs. Three promising aspects of the future of MT are (1) the *automatic learning,* from sets of examples, of enconverters to and deconverters from appropriate *interlingua* representations; (2) the development of authoring tools enabling users to interactively create *self-explaining documents* (SEDs), and to automatically translate these into many other languages; and (3) the *connection of domain ontologies* to MT systems of all types through expert system interfaces. *Ubilingual* (ubiquitous multilingual) computing is already on its way, and progress towards *intelligent* ubilingual computing is anticipated.

References

[1] **Al-Adhaileh M. H., Tang E. K. & Zaharin Y. (2002)** *A synchronization structure of SSTC and its applications in machine translation.* Proc. International Conference on Computational Linguistics (COLING-02) Workshop on Machine translation in Asia, Taiwan, ACL, pp. 1-8.

[2] **Antoine J.-Y. & Genthial D., ed. (1999)** *Actes du colloque Méthodes hybrides TALN/TALP pour le traitement robuste de la langue.* D. Genthial & J.-Y. Antoine ed., Cargèse, France, 146 p.

[3] **Bellynck V. (1998)** *Multimodal Visualization of Geometrical Constructions.* Proc. CVIR'1998, workshop on Content Vizualisation and Intermedia Representations, Montreal, Québec, Canada, August 15, 1998, 6 p.

[4] **Bellynck V., Boitet C. & Kenwright J. (2006)** *Contribute to learn: a proof of concept with ITOLDU.* In "Japio 2006 Year Book", JAPIO (Japan Patent Information Organization), Tokyo, 6 p. (invited chapter, based on a PACLING 2005 paper)

[5] **Bellynck V., Boitet C. & Kenwright J. (2007)** *Bilingual Lexical Data Contributed by Language Teachers Via a Web Service: Quality vs. Quantity.* Proc. CICLING-2007 (Computational Linguistics and Intelligent Text Processing), 19-23 Feb. 2007, UPM, A. Gelbukh ed., 12 p.
[6] **Bellynck V. B., Christian; Kenwright, John (2005)** *ITOLDU, a Web Service to Pool Technical Lexical Terms in a Learning Environment and Contribute to Multilingual Lexical Databases.* Proc. Computational Linguistics and Intelligent Text Processing (Proc. CICLING-2005), Mexico, Mexique, 14-18 Feb., Springer (LNCS 3406), A. Gelbukh ed., pp. 319-327.
[7] **Besacier L., Le V.-B. & Boitet C. (2006)** *ASR and Translation for Under-Resourced Languages.* Proc. ICASSP-06, Toulouse, 15-19/5/06, IEEE, 4 p.
[8] **Bey Y., Kageura K. & Boitet C. (2007)** *BEYTrans: A Free Online Collaborative Wiki-Based CAT Environment Designed for Online Translation Communities.* Proc. PACLIC-21 (Pacific Asia Conference on Language, Information and Computation), Seoul, 1-3 November, 2007, 8 p.
[9] **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, 5-9 Aug. 1994, Y. Wilks ed., vol. 1/2, pp. 115—119.
[10] **Blanchon H. & Boitet C. (2003)** *Two Steps Towards Self-Explaining Documents.* Proc. Convergences-2003, Alexandrie, 2-5/12/2003, 6 p.
[11] **Blanchon H. & Boitet C. (2004)** *Les Documents Auto-Explicatifs : une voie pour offrir l'accès au sens aux lecteurs.* Proc. CIDE-7 (Approches Sémantiques du Document Numérique), La Rochelle, 22-25 juin 2004, pp. 273-290.
[12] **Blanchon H., Boitet C. & Besacier L. (2004)** *Evaluation of Spoken Dialogue Translation Systems: Trends, Results, Problems and Proposals.* Proc. COLING-04, Genève, 23-27/8/04, ACL, 7 p.
[13] **Blanchon H., Boitet C. & Besacier L. (2004)** *Spoken Dialogue Translation Systems Evaluation: Results, New Trends, Problems and Proposals.* Proc. IWSLT-04 (International Workshop on Spoken Language Translation), Nara, September 30 - October 1, ATR, pp. 95-102.
[14] **Blanchon H. & Boitet C. (2006)** *Annotating Documents by Their Intended Meaning to Make Them Self Explaining: An Essential Progress for the Semantic Web.* Proc. FQAS-06 (Flexible Query Answering Systems), Milan, 7-10/6/06, Lecture Notes in Computer Science 4027, Springer, 12 p.
[15] **Blanchon H. & Boitet C. (2008)** *Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche.* Numéro spécial de TAL sur l'évaluation pp 1-33.
[16] **Blanchon H., Boitet C. & Choumane A. (2008)** *Traduction Automatisée Fondée sur le Dialogue et Documents Auto-Explicatifs : bilan du projet LIDIA.* TAL pp 1-25.
[17] **Boitet C. (1976)** *Un essai de réponse à quelques questions théoriques et pratiques liées à la traduction automatique. Définition d'un système prototype.* Thèse d'Etat, Université de Grenoble (USTMG), 250 p.
[18] **Boitet C. & Gerber R. (1984)** *Expert systems and other new techniques in MT.* Proc. COLING-84, 2-6 July 1984, ICCL & ACL, pp. 468-471.
[19] **Boitet C. & Nédobejkine N. (1986)** *Toward integrated dictionaries for M(a)T: motivations and linguistic organization.* Proc. COLING-86, Aug. 1986, IKS, pp. 423—428.
[20] **Boitet C. & Gerber R. (1986)** *Expert Systems and other new techniques in MT.* In "Neue Ansätze in maschineller Sprachübersetzung", Bátori & Weber, Niemeyer, Tübingen, pp. 103—119.
[21] **Boitet C. (1988)** *Software and lingware engineering in modern M(A)T systems.* In "Handbook for Machine Translation", Bátori, ed., Niemeyer.
[22] **Boitet C. (1988)** *Bernard VAUQUOIS' Contribution to the Theory and Practice of building MT Systems : a historical perspective.* Proc. Second International Conference on Theoretical and Methodological Issues in the Machine Translation of Natural Languages, June 1988, CMT, CMU, J. Carbonell ed., 18 p.
[23] **Boitet C. (1988)** *PROs and CONs of the pivot and transfer approaches in multilingual Machine Translation.* Proc. Int. Conf. on "New directions in Machine Translation", 18–19 August 1988, Foris Publications, pp. 93—108.
[24] **Boitet C. (1990)** *Towards Personal MT : general design, dialogue structure, potential role of speech.* Proc. COLING-90, 20-25 août 1990, ACL, vol. 3/3, pp. 30-35.
[25] **Boitet C. (1990)** *Multilingual Machine Translation does not have to be saved by Interlingua.* Proc. MMT'90, Tokyo, 5-6 Nov. 1990, 2 p.
[26] **Boitet C. (1991)** *Twelve Problems for Machine Translation.* Proc. CICL-91 (International Conference on Current Issues in Computational Linguistics), USM, Penang, June 1991, 12 p.
[27] **Boitet C. (1992)** *Translation workstations in the year 2000.* Proc. Int. Symposium on Machine Translation, 28-30 sept. 1992, 11 p.
[28] **Boitet C. (1993)** *La TAO comme technologie scientifique : le cas de la TA fondée sur le dialogue.* In "La traductique", A. Clas & P. Bouillon, ed., Presses de l'Université de Montréal, Montréal, pp. 109—148.
[29] **Boitet C. (1993)** *TA et TAO à Grenoble… 32 ans déjà !* (T.A.L. (revue semestrielle de l'ATALA)), **33**/1—2, Spécial Trentenaire, pp 45—84.
[30] **Boitet C. (1993)** *Practical Speech Translation Systems will integrate human expertise, multimodal communication, and interactive disambiguation.* Proc. MTS-IV, 18–22 July 1993, pp. 173–176.
[31] **Boitet C. & Seligman M. (1994)** *The "Whiteboard" Architecture: a way to integrate heterogeneous components of NLP systems.* Proc. COLING-94, 5—9 August 1994, pp. 243—246, 7 p.
[32] **Boitet C. (1994)** *Dialogue-Based MT and self-explaining documents as an alternative to MAHT and MT of controlled languages.* Proc. Machine Translation 10 Years On, 11-14 Nov. 1994, Cranfield University Press, D. Clarke & A. Vella ed., pp. 22.1—9.
[33] **Boitet C. & Blanchon H. (1994)** *Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup.* Machine Translation **9**/2, pp 99—132.
[34] **Boitet C. (1995)** *Factors for success (and failure) in Machine Translation — some lessons of the first 50 years of R&D.* Proc. MTS-V (Fifth Machine Translation Summit), 11—13 July 1995, CEE, Y. Wilks ed., 17 p.

[35] **Boitet C. (2001)** *Méthodes d'acquisition lexicale en TAO : des dictionnaires spécialisés propriétaires aux bases lexicales généralistes et ouvertes.* Proc. Conférence associée à TALN-2001 "TAO et applications de grande taille", Tours, 5 juillet 2001, ATALA, D. Maurel ed., vol. 2/2, pp. 249-265.
[36] **Boitet C. (2001)** *Four technical and organizational keys for handling more languages and improving quality (on demand) in MT.* Proc. MTS2001 Workshop on "MT2010 — Towards a Road Map for MT", Santiago de Compostela, 18/9/01, IAMT, 8 p.
[37] **Boitet C. (2001)** *Machine Translation.* In "Encyclopedia of Cognitive Science", Nature Publishing Group, London, (in manuscript form) 24 p.
[38] **Boitet C. (2002)** *A rationale for using UNL as an interlingua and more in various domains.* Proc. LREC-02 First International Workshop on UNL, other Interlinguas, and their Applications, Las Palmas, 26-31/5/2002, ELRA/ELDA, J. Cardeñosa ed., pp. 23—26.
[39] **Boitet C. & Tsai W.-J. (2002)** *La coédition langue<—>UNL pour partager la révision entre les langues d'un document multilingue : un concept unificateur.* Proc. TALN-02, Nancy, 22-26/6/2002, ATALA, J.-M. Pierrel ed., 12 p.
[40] **Boitet C. & Tsai W.-J. (2002)** *Coedition to share text revision across languages.* Proc. COLING-02 WS on MT, Taipeh, 1/9/2002, 8 p.
[41] **Boitet C. (2002)** *A roadmap for MT.* In "Festschrift für Winfried Lenders", Bonn, 7 p.
[42] **Boitet C. (2003)** *Automated Translation / Traduction Automatique.* Revue française de linguistique appliquée VIII-2, pp 99-121.
[43] **Boitet C. (2005)** *Gradable quality translations through mutualization of human translation and revision, and UNL-based MT and coedition.* In "Universal Networking Language, advances in theory and applications", J. Cardeñosa, A. Gelbukh & E. Tovar, ed., Research in Computing Science, IPN-CIC, Mexico, pp. 393—410. (presented at the 2nd Workshop on UNL and Other Interlinguas)
[44] **Boitet C. (2005)** *New architectures for "democratic" tunable quality MT systems.* Proc. Pacific Association for Computational Linguistics (Proc. PACLING-2005), Meisei daigaku, Hino campus, 24-27 Aug. 2005, PACLING Association, H. Sakaki ed., pp. 33—57. (invited communication)
[45] **Boitet C., Bey Y. & Kageura K. (2005)** *Main research issues in building web services for mutualized, non-commercial translation.* Proc. SNLP-05, Bangkok, 12-16/12/05, Chulalongkorn university, 13 p.
[46] **Boitet C., Bey Y., Tomokiyo M., Cao W. & Blanchon H. (2006)** *IWSLT-06: Experiments with Commercial MT Systems and Lessons from Subjective Evaluations.* Proc. IWSLT-06 (International Workshop on Spoken Language Translation), Kyoto, 27-28/11/06, ATR, pp. 23—30.
[47] **Boitet C. (2007)** *Corpus pour la TA : types, tailles, et problèmes associés, selon leur usage et le type de système.* RFLA (Revue Française de Linguistique Appliquée) Paris, **1**/pp 25-38.
[48] **Boitet C., Boguslavskij I. & Cardeñosa J. (2007)** *An Evaluation of UNL Usability for High Quality Multilingualization and Projections for a Future UNL++ Language.* In "Computational Linguistics and Intelligent Text Processing (Proc. CICLING-2007)", A. Gelbukh, ed., Springer (LNCS 4394), pp. 361-373. (ISBN-10: 3-540-70938-X Springer, ISSN: 0302-9743)
[49] **Boitet C. (2007)** *Linguistic & computational MT architectures are orthogonal: elements for choice.* Proc. SNLP-07, Pattaya, 15/12/07, Kasetsart University, A. Kawtrakul ed., 81 p. (slides, keynote presentation)
[50] **Boitet C. (2008)** *Les architectures linguistiques et computationnelles en traduction automatique sont indépendantes.* Proc. TALN-08, Avignon, 9-13/6/08, ATALA ed., 12 p.
[51] **Boitet C. (2009)** *Machine Translation (MT) and Computer-Aided Translation (CAT) -- Lecture 1: Linguistic, computational and operational architectures of MT systems.* NII lecture series, Tokyo, 14 April 2009, NII, 44 p. (slides available on NII web site)
[52] **Boitet C. (2009)** *Machine Translation (MT) and Computer-Aided Translation (CAT) -- Lecture 2: Linguistic architectures of MT systems.* NII lecture series, Tokyo, 16 April 2009, NII, 64 p. (slides available on NII web site)
[53] **Boitet C. (2009)** *Machine Translation (MT) and Computer-Aided Translation (CAT) -- Lecture 3: Computational architectures of MT systems.* NII lecture series, Tokyo, 21 April 2009, NII, 35 p. (slides available on NII web site)
[54] **Boitet C. (2009)** *Machine Translation (MT) and Computer-Aided Translation (CAT) -- Lecture 4: Engineering of MT and CAT systems (CAT = MT + TA, TA = translation aids).* NII lecture series, Tokyo, 23 April 2009, NII, 25 p. (slides available on NII web site)
[55] **Boitet C. (2009)** *Machine Translation (MT) and Computer-Aided Translation (CAT) -- Lecture 5: Evaluation of MT and CAT systems for various operational architectures.* NII lecture series, Tokyo, 28 April 2009, NII, 22 p. (slides available on NII web site)
[56] **Boitet C. (2009)** *Machine Translation (MT) and Computer-Aided Translation (CAT) -- Lecture 6: Corpora for hybrid MT/CAT systems.* NII lecture series, Tokyo, 30 April 2009, NII, 21 p. (slides available on NII web site)
[57] **Boitet C., Huynh C.-P., Blanchon H. & Nguyen H.-T. (2009)** *A Web-oriented System to Manage the Translation of an Online Encyclopedia Using Classical MT and Deconversion from UNL.* Proc. RIVF 2009, Da Nang, 13-17/7/09, IEEE, P. Bellot ed., 8 p. (extended & updated from eponym CI-2008 paper)
[58] **Brunet-Manquat F. (2004)** *Syntactic parser combination for improved dependency analysis.* Proc. ROMAND-2004, Workshop COLING 2004, 24-28/8/04, vol. 1/1, 7 p.
[59] **Callison-Burch C., Osborne M. & Koehn P. (2006)** *Re-evaluating the Role of BLEU in Machine Translation Research.* Proc. EACL-06, Trento, April 3-7, 2006, ITC/irst ed., 8 p.
[60] **Chandioux J. & Guérard M.-F. (1981)** *METEO: un système à l'épreuve du temps.* META (Journal des Traducteurs) 1, pp 17—22.
[61] **Choumane A. (2005)** *Accès à des services de traduction interactive fondée sur le dialogue à travers un éditeur XML.* Mémoire de Master 2 en informatique (M2R), GETA/CLIPS et INRIA, UJF, 76 p.
[62] **Do T.-N.-D., Le V.-B., Bigi B., Besacier L. & Castelli E. (2009)** *Exploitation d'un corpus bilingue comparable pour la création d'un système de traduction probabiliste vietnamien-français.* Proc. TALN-09, Senlis, 23-26/6/09, pp. 10.
[63] **Dolan W. B., Pinkham J. & Richardson S. D. (2002)** *The Microsoft Research Machine Translation System.* In "Machine Translation: From Research to Real Users", Lecture Notes in Computer Science 2499/2002, Springer Berlin / Heidelberg, pp. 237-239.
[64] **Fafiotte G., Boitet C., Seligman M. & Zong C. (2004)** *Collecting Bilingual Dialogues using a Web-Based Platform for the Study of Interpretation.* Proc. LREC-04 (Language Resources and Evaluation Conference), Lisbonne, 24-28/5/04, 9 p.

[65] **Gallup S. (1993)** *Caterpillar Technical English and Automatic Machine Translation.* Proc. 40th Annual Conference of the Society for Technical Communication (STC-93), Dallas, June 6-10, 1993, pp. 421-424. http://www.stc.org/ConfProceed/1993/PDFs/Pg421424.pdf
[66] **Gao Y., Gu L., Zhou B., Sarikaya R., Afify M., Kuo H.-K. & al. (2006)** *IBM MASTOR System: Multilingual Automatic Speech-to-speech Translator.* Proc. HLT-NAACL 2006, 4 p.
[67] **Gerber R. & Boitet C. (1985)** *On the design of expert systems grafted on MT systems.* Proc. TMI-85 (Conf. on Theoretical and Methodological Issues in Machine Translation of natural language), Aug. 1985, Colgate Univ., pp. 116-134.
[68] **Hajlaoui N. & Boitet C. (2008)** *TA statistique à petits corpus pour de petits sous-langages.* Proc. TOTh-08, Avignon, 5-6/6/08, Université de Savoie, L. Damas ed., 18 p.
[69] **Hovy E., King M. & Popescu-Belis A. (2002)** *Principles of Context-Based Machine Translation Evaluation.* Machine Translation (Machine Translation), **17**/1, pp 43-75.
[70] **Huang X. M. (1990)** *A Machine Translation System for the Target Language Inexpert.* Proc. COLING-90, 20-25 Aug. 1990, ACL, vol. 3/3, pp. 364-367.
[71] **Hutchins W. J. (1986)** *Machine Translation : Past, Present, Future.* Ellis Horwood, John Wiley & Sons, Chichester, England, 382 p.
[72] **Hutchins W. J. & Somers H. L. (1992)** *An Introduction to Machine Translation.* H. B. Jovanovich, ed., Academic Press, 362 p.
[73] **Huynh C.-P., Boitet C. & Blanchon H. (2008)** *SECTra_w : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora.* Proc. LREC-08, Marrakech, 27-31/5/08, ELRA/ELDA ed., 8 p.
[74] **Johnson D. E. & Watanabe H. (1991)** *Relational-grammar-based generation in the JETS Japanese-English machine translation system.* Machine Translation Collection Humanities, Social Sciences and Law, **6**/1, pp 1-20.
[75] **Lafourcade M. & Boitet C. (2002)** *UNL Lexical selection with conceptual vectors.* Proc. LREC-02 (International Conference on Language Resources and Evaluation), Las Palmas, 26-31/5/2002, A. Zampolli & I. Prodanoff ed., vol. VI/VI, pp. 1958—1964.
[76] **Langlais P., Thomas Leplus S. G. & Lapalme G. (2005)** *From the RealWorld to Real Words: the METEO case.* Proc. EAMT-05, Budapest, May 2005, 10 p.
[77] **Melby A. K., Smith M. R. & Peterson J. (1980)** *ITS : An Interactive Translation System.* Proc. COLING-80, Tokyo, 30/9-4/10/80, pp. 424—429.
[78] **Melby A. K. (1981)** *Translators and Machines - Can they cooperate ?* (META), **26**/1, pp 23-34.
[79] **Melby A. K. (1982)** *Multi-Level Translation Aids in a Distributed System.* Proc. COLING-82, 5-10 July 1982, vol. 1/2, pp. 215-220.
[80] **Melby A. K. (1991)** *MicroMATER. A Proposed Standard Format for Exchanging Lexical / Terminological Data Files.* Meta : journal des traducteurs / Meta: Translators' Journal "La terminologie dans le monde : orientations et recherches." Sous la direction de Monique C. Cormier et Jacques Lethuillier, **36**/pp 135-160.
[81] **Nguyen H.-T., Boitet C. & Sérasset G. (2007)** *PIVAX, an online contributive lexical database for heterogeneous MT systems using a lexical pivot.* Proc. SNLP-07, Pattaya, 13-16/12/07, Kasetsart University, A. Kawtrakul ed., 6 p.
[82] **Nirenburg S. & al. (1989)** *KBMT-89 Project Report.* Center for Machine Translation, Carnegie Mellon University, Pittsburg, April 1989, pp. 286.
[83] **Nirenburg S. (1989)** *Knowledge-based Machine Translation.* (Machine Translation), 4, pp 5-24.
[84] **Nyberg E., Kamprath C. & Mitamura T. (1998)** *The KANT Translation System From R&D to Large-Scale Deployment.* Localization Industry Standards Association (LISA) Newsletter **2**/pp 1-8.
[85] **Philipp Koehn H. H., Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst (2007)** *Moses: Open Source Toolkit for Statistical Machine Translation,.* Proc. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.
[86] **Satayamas V., Boitet C. & Kawtrakul A. (2008)** *Specialized editor for annotating word boundaries collaboratively.* Proc. IJCNLP-08 workshop on NLP of Indian languages, Hyederabad, 7-12/1/08, P. Battacharyya ed., 8 p.
[87] **Seligman M. & Boitet C. (1993)** *A "Whiteboard" Architecture for Machine Interpretation.* Proc. International Symposium on Spoken Dialogue, 10–12 November 1993, Waseda University, 4 p.
[88] **Seligman M. & Dillinger M. (2008)** *Rapid Portability among Domains in an Interactive Spoken Language Translation System.* Proc. COLING-2008 workshop on Speech Processing for Safety Critical Translation and Pervasive Applications, Manchester, UK, 24 Aug. 2008, pp. 40-47. http://www.aclweb.org/anthology/W08-1508
[89] **Uchida H. (2004)** *The Universal Networking Language (UNL) Specifications Version 3 Edition 3.* UNL Center, UNDL Foundation, December 2004, 43 p. http://www.undl.org/unlsys/unl/UNLSpecs33.pdf
[90] **Uchino H., Shirai S., Yokoo A., Oyama Y. & Furuse O. (2001)** *ALTFLASH: A Japanese-English Machine Translation System for Market Flash Reports.* EICE Transactions on Information and Systems, Pt.2 (Japanese Edition) **VOL.J84-D-2**/6, pp 1167-1174.
[91] **Vauquois B. & Chappuy S. (1985)** *Static grammars: a formalism for the description of linguistic models.* Proc. TMI-85 (Conf. on theoretical and metholodogical issues in the Machine Translation of natural languages), Colgate Univ., Hamilton, N.Y., Aug. 1985, pp. 298-322.
[92] **Zaharin Y. (1990)** *Generation of Synthesis Programs in ROBRA (ARIANE) from String-Tree Correspondence Grammars.* Proc. COLING-90, Helsinki, 18—25 Aug. 1990, pp. 425—430.
[93] **Zaharin Y. & Tang E. K. (1991)** *String-Tree Correspondence Grammars as a base gor the automatic generation of analysis programs in machine tranaslation. Proceedings of the.* Proc. International Conference on Current Issues in Computational Linguistics, Penang, June 1991, 10 p.
[94] **Zhou B., Gao Y., Sorensen J., Déchelotte D. & Picheny M. (2003)** *A Hand-Held Speech-To-Speech Translation System.* Proc. HLT-NAACL, Edmonton, Canada, May-June 2003, 6 p.