The Contribution of End-Users to the TransType2 Project

Elliott Macklovitch[1]

RALI Laboratory, Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Canada H3C 3J7
macklovi@iro.umontreal.ca

Abstract. TransType2 is a novel kind of interactive MT in which the system and the user collaborate in drafting a target text, the system's contribution taking the form of predictions that extend what the translator has already typed in. TT2 is also an international research project in which end-users are represented by two translation firms. We describe the contribution of these translators to the project, from their input to the system's functional specifications to their participation in quarterly user trials. We also present the results of the latest round of user trials.

1 Introduction

The goal of the TransType2 project (Foster et al. 2002) is to develop a novel type of *interactive* machine translation system. The system observes the user as s/he types a translation, attempts to infer the target text the user has in mind and periodically proposes extensions to the prefix which the user has already keyed in. The user is free to accept these completions, modify them as desired or ignore them by simply continuing to type. With each new character the user enters, the system revises its predictions in order to make them compatible with the user's input.

In itself, interactive machine translation (IMT) is certainly not novel; in fact, the first attempts at IMT go back to the MIND system, which was developed by Martin Kay and Ron Kaplan at the RAND Corporation in the late 1960's.[2] There have been numerous subsequent attempts to implement IMT, some of which gave rise to commercial systems, like ALPS' ITS system, while others have been embedded in controlled language systems, like the KANT system developed at CMU (Nyberg et al. 1997).[3] What all of these previous efforts share in common is that the focus of the interaction between the user and the system is on the *source text*. In particular, whenever the system is unable to disambiguate a portion of the source text, it requests assistance from the user. This can be to help resolve various types of source language ambiguity, such as the correct morpho-syntactic category of a particular word, syntactic dependencies between phrases, or the referent of an anaphor. In principle, once the user has provided the system with the

[1] The work described in this article is the fruit of a sustained collaborative effort, and I want to express my gratitude to all the participants in the TT2 Consortium, particularly to the translators who are testing successive versions of the system.

[2] For more on MIND, see (Hutchins 1986), pp.296-297.

[3] Other IMT systems specifically focus on multi-target translation; see for example (Blanchon and Boitet 2004) and (Wehrli 1993).

information necessary to disambiguate the source text, the system can then complete its analysis and continue to properly translate the text into the target language.

This is not the place to enumerate all the difficulties that have dogged this classic approach to interactive MT; however, there are a few important differences with TransType which we should point out. Suppose that the user of the system is a translator, as was often the case in the early decades of MT. Notice that the kind of information being solicited from the user by these classic IMT systems does not focus on translation knowledge per se, but instead involves formal linguistic analysis, of a kind that many translators have not been trained to perform. In contrast, the focus of the interaction between the user and the system in TransType is squarely on the drafting of the target text. After reading the current source text segment, the translator begins to type his/her desired translation. Based on its analysis of the same source segment and using its statistical translation and language models, TransType immediately proposes an extension to the characters the user has keyed in. The user may accept all or part of the proposed completion, or s/he may simply go on typing; in which case, the system continues trying to predict the target text the user has in mind. When the system performs well, the user will normally accept these machine-generated completions, thereby diminishing the number of characters s/he has to type and hopefully reducing overall translation time. But the important point is that in this paradigm both the user and the system contribute in turn to the drafting of the *target* text, and the translator is not solicited for information in an area in which s/he is not an expert.

Another important difference between classic IMT systems and the target-text mediated approach of TransType may be formulated in this way: Who leads? In the classic IMT approach, it is the system that has the initiative in the translation process; the system decides when and what to ask of the user, and once it has obtained the required information from the user, the system will autonomously generate its translation, much like any other fully automatic MT system. In the best of circumstances, the system will succeed in producing a grammatical and idiomatic sentence in the target language which correctly preserves the meaning of the source sentence. But even in this ideal situation, it would mistaken to believe that this is the *only* correct translation of the source sentence; for as every translator knows, almost any source text admits of multiple, equally acceptable target language renditions. As (King et al. 2003) put it:

> "There is no one right translation of even a banal text, and even the same translator will sometimes change his mind about what translation he prefers." (p.227)

What happens if the translation generated by the system does not correspond to the one which the user had in mind? One of two things: either the user changes his/her mind and accepts the machine's proposal; or the user post-edits the system's output, changing it so that it conforms to the translation s/he intended. But in either case, it is the user who is responding to, or following the system's lead. In TransType, on the other hand, it is entirely the other way round. The user guides the system by providing prompts in the form of target text input, and the system reacts to those prompts by trying to predict the rest of the translation which the user is aiming for. Moreover, the system must *adapt* its predictions to changes in the user's input. Here, quite clearly, it is the user who leads and the system which must follow.

This target-text mediated interactive MT is certainly a intriguing idea – but will it work? Only the system's intended end-users, i.e. professional translators, can answer that question. The TransType2 (henceforth TT2) Consortium includes two translation firms, one in Canada (Société Gamma Inc.) and one in Spain (Celer Soluciones S.L.). These partners play a very important role in the TT2 project, serving to balance its ambitious research program with the concrete needs of real end-users. The project provides for

various channels through which the end-users can interact with the research teams who are developing the translation engines. One of the most important of these are the user trials that begin about half-way through the project and continue right up to its conclusion, at month thirty-six.

In the following section, we describe in more detail the role of these end-users in the TransType2 project. In section 3, we present the protocol for the latest round of user evaluations, which have just been completed at Société Gamma and at Celer Soluciones. In section 4, we report on the main results obtained in those trials – results which are necessarily tentative, since the project still has more than a year to run. In the final section, we draw some conclusions about the future of IMT.

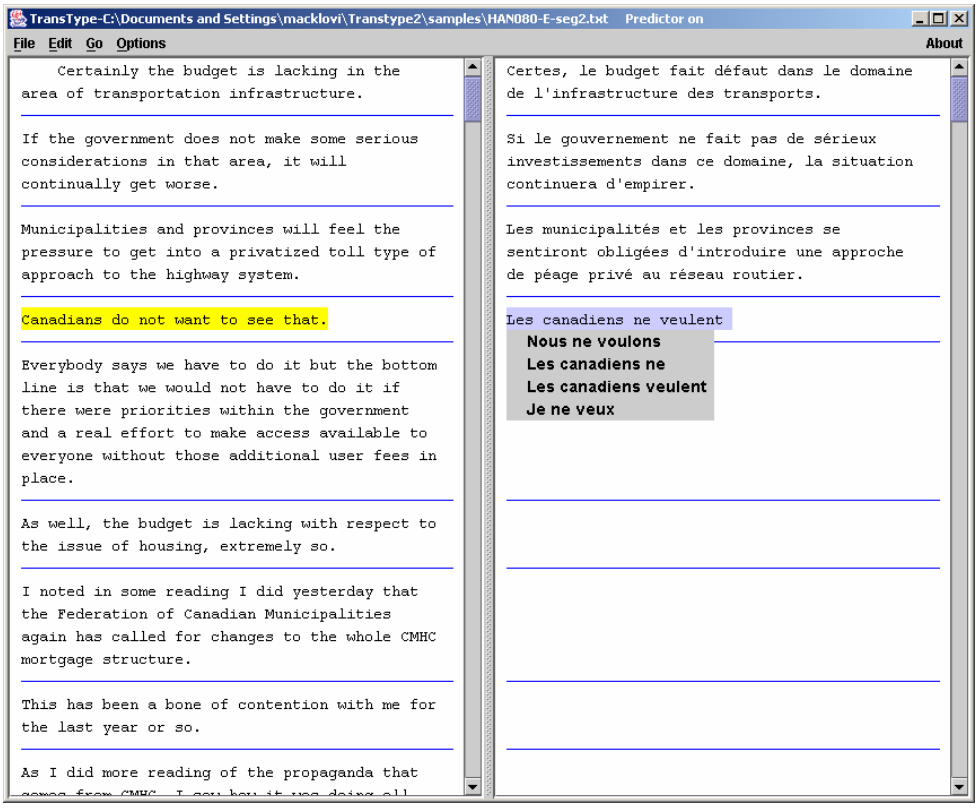

Fig. 1. Snapshot of a TransType session

2 The Role of End-Users in the TT2 Project

TransType2 is a three-year research project that was officially launched in March 2002. In Europe, it is funded under the EC's Fifth Framework Program, and in Canada by both the federal government (through NSERC) and by the Quebec government's Department of Research, Science and Technology. One of the basic goals of TT2 is to provide a framework in which leading-edge research can be conducted in the area of data-driven methods in NLP and, more specifically, in machine translation. But TT2 also has another

major objective, which is to provide a practical application for that research which will hopefully help solve a pressing social problem, to wit: how to meet the ever-growing demand for high-quality translation. Our research in TT2 is aimed at developing an innovative machine-aided translation system which should facilitate the task of producing high-quality translations and make that task more cost-effective for human translators.

As mentioned above, the TT2 Consortium includes two translation firms, and the participating translators at Société Gamma and Celer Soluciones have been directly and actively involved in the project right from the outset. In particular, the translators played a key role in the drafting of the system's functional specifications and in helping to design its graphical user interface (GUI). As an example of their input on the functional specs, the translators insisted on the fact that the system needs to be open to terminological input from the user, in the form of glossaries that would contain client- or domain-specific terminology. Any system that could not be customized in this way, they told us, would be seriously handicapped, because it would force the user to repeatedly correct terminology in the target text that was available in his/her glossary. Such a requirement may not be problematic in rule-based MT systems, where the user can generally add a specialized glossary or modify the content of the system's dictionaries, these being formalized in a manner that is more or less transparent to a human. This is not the case, however, in most statistical MT systems, where there is no user interface to a distinct lexical component. Hence, the translators' requirement raised an interesting research problem for the engine developers on the TT2 project: first, how to make these declarative, user-supplied glossaries compatible with the system's statistical translation engine, and then how to grant priority to the entries in the glossary over the translations previously inferred by the system from its training corpus.

As a (somewhat paradoxical) example of user input on the specs for TT2's GUI, the translators told us that there must be an easy way to shut off the system's predictions. In particular, they found during pre-trial testing that when they went back into a completed segment to lightly revise or correct the translation, the system's predictions cluttered up the screen and proved more of a hindrance than a help. As a consequence, the team that was developing the GUI added a number of new options to the interface. One is called "never-within-text" and, as its name suggests, it blocks the display of system completions whenever there is text to the right of the cursor. Another option that was added to the GUI is a delay setting which allows the user to specify a certain interval of inactivity, e.g. 3 seconds, during which the system displays no predictions. If the user stops typing for more than the specified interval, as when s/he is searching for a solution to a particular translation problem, only then does the system display its predictions. Of course, the user can always summon up a completion by hitting a keyboard short-cut, even when the prediction engine is turned off.

The other major contribution of the users in the TT2 project involves their participation in the quarterly trials that began in month eighteen. These are intended to evaluate the *usability* of successive system prototypes and, in this respect, are quite different from the internal technology evaluations, which also form part of the project work plan. The aim of the latter is to gauge progress in the core technology, i.e. improvements in the statistical translation engines; and to do this, the principal means employed are automatic metrics such as word error rate or methods like BLEU. The usability evaluation, on the other hand, inescapably involves the intended end-users of TransType, i.e. professional translators; and here, the goal is to evaluate, not so much the performance of the system in vitro (as it were), but its actual impact on the productivity of working translators and the ease (or difficulty) with which they adapt to the system. An equally important objective of the user trials is to provide a channel of communication through which the participating translators

can furnish feedback and suggestions to the system developers, so that the latter can continue to make improvements to the system.[4]

We have just completed the third round of user evaluations in the TT2 project, and the first in which the participating translators at Gamma and at Celer have actually had the opportunity to work with the system in a mode that approximates their real working conditions. In the following section, we present our objectives for this round of user trials and the protocol which governed its organization.

3 The Protocol for Evaluation Round 3 (ER3)

The corpus selected for ER3 came from a Xerox User Guide for a large commercial printer; it is part of an approximately one million word collection provided by XRCE, one of the partners in the TT2 consortium. Each of the labs developing prediction engines in the project – RWTH in Aachen, ITI in Valencia, and RALI in Montreal – trained its system on this Xerox corpus, excluding of course the chapters that were to be translated during the user trial. For our test corpus, we decided to use the same chapters at the two translation firms. At Celer in Madrid, these would be translated from English into Spanish; at Gamma in Ottawa, they would be translated from English into French. Because these manuals are relatively technical in nature, XRCE also provided a terminology glossary with about 750 entries, as well as a PDF version of the original English manual containing tables and graphics.[5]

One of our sub-objectives for ER3 was to try to determine if users had a marked preference for a single lengthy completion versus multiple predictions (which may or may not be shorter than the single prediction, depending on the engine). To this end, the participants at each site were asked to translate chapters from the Xerox printer manual using two different prediction engines. At Gamma, the participants would translate two test chapters into French with the RWTH engine, configured to provide single full-sentence length predictions, without alternate choices; and they would translate two other chapters with the RALI engine, configured to provide five alternate predictions of shorter length.[6] At Celer, the participants would also use the RWTH engine, configured to provide single full-sentence completions; and instead of the RALI's engine, they would use ITI's engine, configured to provide multiple completions of varying length. All three engines were embedded within the same GUI, which is shown in Figure 1 above.

In order to obtain a baseline comparison of their productivity on this kind of technical material, we also asked the translators to translate one chapter of the Xerox manual using TT2, but with the prediction engine turned off. As can be seen from the snapshot in Figure 1, TransType's GUI takes the form of a two-paned text editor in which the source text appears in the left-hand pane, divided into sentence-like segments. As soon as the user selects a given segment, the system responds by inserting a proposed translation for it in the corresponding cell in the right-hand pane. In this first "dry-run", the prediction engine

[4] To encourage users to provide such feedback, TT2 includes a pop-up notepad, with entries that are automatically time-stamped and identified with the user's name.

[5] Currently, TT2 only accepts plain text files as input. By consulting the PDF original, the participants could situate certain segments extracted from tables or graphics within their proper context.

[6] At the time of the ER3 trial, the RALI's maximum entropy engine could not provide completions longer than five words. ITI's engine, which was the other system configured to provide multiple predictions, was able to provide longer predictions, up to full-sentence length.

was turned off and the users simply typed their own translation in the right-hand pane. What isn't evident from the snapshot in Figure 1 is that the GUI also generates a detailed trace file, which records every one of the user's actions and all the system's predictions, including their precise times. To determine each participant's baseline productivity, we had only to consult the trace file, determine how many words the user had translated and divide it by the time expired.

In the remaining four half-day sessions of this evaluation round, the system's prediction engine was turned back on and the participants were asked not to modify the interface parameters, in order to facilitate the comparison of their results.

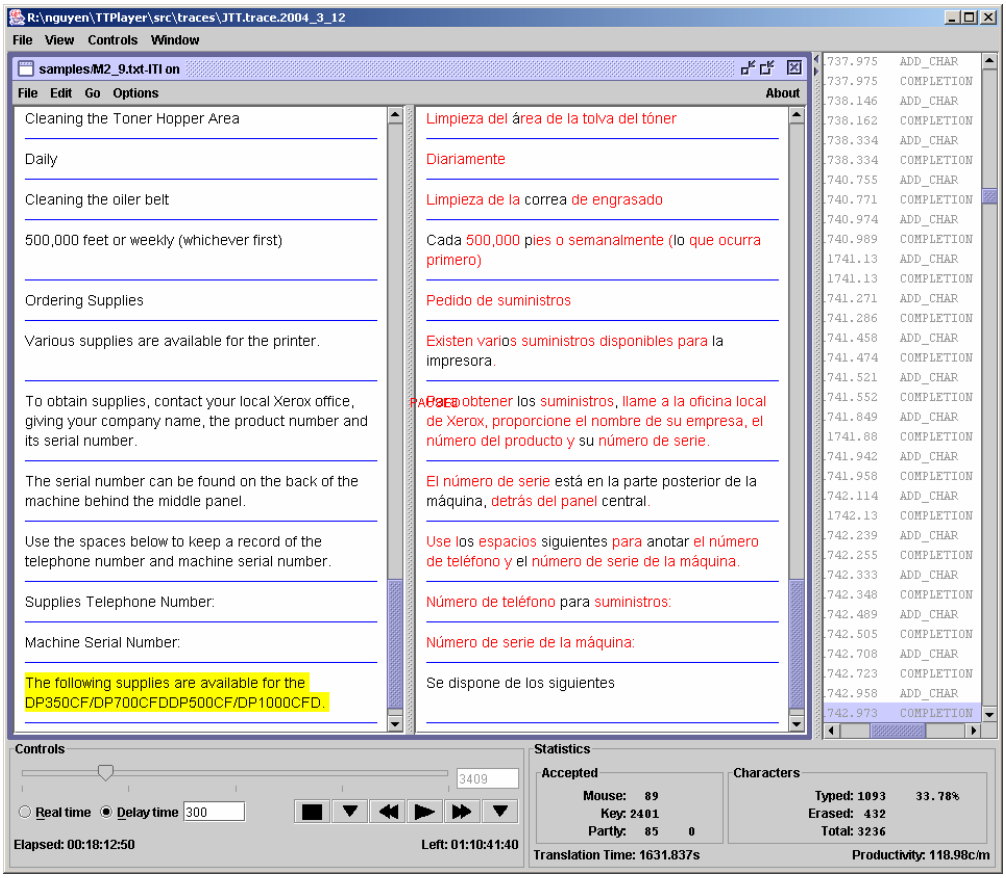

Fig. 2. Snapshot of TT-Player

4 TT-Player and the Analysis of the Trace Files

We mentioned in the previous section that every interaction between the user and the system is recorded in the trace file, making these files very detailed, lengthy and rather onerous to consult by hand. In order to facilitate the evaluation process, the RALI has developed a utility program called TT-Player which is designed to read the trace file of a TT2 session and play it back, much like a VCR player plays back a video tape. Moreover,

TT-Player also produces a statistical summary of the session, highlighting whatever statistics we wish to bring out.

Figure 2 above contains a snapshot of TT-Player taken in replay mode. In this session, the English source text appears in the left-hand pane and, in the middle pane of the main window, the Spanish translation in progress. The narrow pane on the right contains the trace, showing both the actions performed by the user and the completions proposed by the system, including their precise times. This is what is being automatically played back and, like a VCR, the replay can be controlled via the arrow buttons on the bottom of the main frame. In the translation pane, sequences typed by the user appear in black, while those that derive from system predictions appear in red (or are pale if this page isn't printed in colour). At the bottom right of the main frame, certain statistics are provided which are automatically updated with each action. A complete statistical analysis of the session is also available via the View menu.

Given a trace file of this detail, it is possible to extract a broad variety of measurements and statistical indicators from the raw data. In the TT2 research project, there are basically two things that we want to measure: first, the impact that our IMT system has on translators' productivity; and second, the manner in which the translators actually make of use the system, i.e. which features they take advantage of and which they ignore. The latter kind of data should help the developers improve the design of the system, while the former should inform us of the general viability of this novel approach to IMT. Although the NLP literature is replete with methodologies for evaluating MT systems, the great majority of these have been designed for fully automatic MT systems and hence are not entirely appropriate for an interactive system like TransType.[7] The ISLE project developed a particularly thorough and rigorous MT evaluation taxonomy, which is now available online.[8] Here is what we find there on metrics for interactive translation:

> Metric: Steps for translation – method: Count the number of times system requires assistance when translating a test corpus. – Measurement: Number of steps needed or number of steps as percentage of test corpus size.
>
> Metric: Time for interactive translation – Method: Measure the amount of time it takes to perform interactive translation on test corpus. – Measurement: Amount of time for interactive translation on test corpus.

The first metric, notice, betrays a certain bias toward classic IMT systems; the tacit assumption seems to be that the fewer times the system requires the user's assistance, the better. Our bias in target-text mediated IMT is quite different. What we want to count is the number of times that the user accepts the system's proposals in drafting his/her translation; and in principle, the more often s/he does this, the better. As for the second FEMTI metric, this is precisely the way we have adopted to measure our participants' productivity. The following table lists the parameters that TT-Player was programmed to extract from the trace files on ER3. It also summarizes the results of one of the participants on the "dry-run", when the prediction engine was turned off, and on a second session, with one of the two prediction engines turned on.

---

[7] In the context of its work on the original TransType project, the RALI did elaborate an evaluation methodology specifically designed for interactive MT; see (Langlais et al. 2002). Needless to say, we drew heavily on this experience.

[8] C.f. http:// www.issco.unige.ch/projects/isle/femti/

G-TR3 (English to French)	**Chapter M2_4** ("dry-run")	**Chapter M2_2** (RWTH Engine)
date of translation	March 8, 2004	March 10, 2004
# words / segments in source	1170 w / 115 seg	3514 w / 278 seg
# words in target / # segments translated	867 w / 81 seg	2773 w / 192 seg
translation time	60.1 min	147 min
Productivity	**14.4 w/min**	**18.9 w/min**
# system predictions		5961
% accepted predictions		6.4 %
average length of accepted predictions		4.3 w
average time to accept a completion		11.4 sec
# entire predictions accepted via keyboard		95
# partial completions accepted via keyboard		22
# entire completions accepted via mouse		0
# partial completions accepted via mouse		267
% of first completions accepted		N/A*
% of entire first completions accepted		25%
Ratio: words accepted / words in target		0.59
Ratio: chars typed / chars in target		0.52
Ratio: deleted chars / chars in target		0.09
Ratio: mouse & kb actions per target word	8.6 actions / w	3.8 actions / w

Table 1. Partial results of one participant on ER3

From the table, we see that by using the predictions provided by the RWTH engine, the translator was actually been able to increase her productivity from 14.4 words per minute on the dry-run to an impressive 18.9 words per minute on Chapter M2_2. During this two and a half hour session, the system proposed 5961 completions, of which the translator accepted 6.4%.[9] The average length of an accepted completion was 4.3 words, and the average time required to accept a completion was 11.4 seconds. The next six lines provide information on the manner in which the user accepted the system's proposals: in whole or in part, using the keyboard or the mouse. The final four lines furnish various ratios between the length of the participant's target text (in words or in characters) and different types of actions, e.g. the number of characters typed or deleted during the session. In the final line, we see that translating Chapter 2_4 on her own, the translator required an average of 8.6 keystrokes or mouse-clicks per word, whereas on Chapter M2_2, with the benefit of the system's predictions, the number of actions per word dropped to 3.8.

[9] This number of predictions may appear at first to be very high; but then it must be remembered that TransType revises its predictions with each new character the user enters.

5 The Results

Before presenting a synthesis of the results we obtained on ER3, a number of caveats are definitely in order. As we mentioned above, this was the first time that the participants at Gamma and at Celer were actually translating with TT2 in a mode that resembles their real working conditions; but "resembles" is the operative word here. TransType remains a research prototype and as such its editing environment does not offer all the facilities of a commercial word processor, e.g. automatic spell checking or search-and-replace. Moreover, this was a very small-scale test, involving only four texts and less than ten thousand words of translation. Hence, the results we present below must be viewed as tentative. At least two other evaluation rounds are planned before the end of the TT2 project, during which the participants will be asked to work with the system for longer periods. Finally, there is another important caveat which should cause us to be cautious, and it has to do with quality controls, of which there were none in this round. During the preparatory sessions at the two agencies, the participants were asked to produce translations of "deliverable" quality; but in fact, we did nothing to ensure that this was the case, relying only on the translators' sense of professionalism. Hence, there was nothing to prevent one participant from rushing through his/her translation, without attempting to reread or polish it, while another might well invest significant time and effort in improving the quality of the final text, even though this would have a negative impact on his/her productivity.

With these caveats in mind, let us now turn to the "bottom-line" quantitative results on ER3. Assuming that the baseline figures provided on the dry-run are reliable, three of the four participants succeeded in increasing their productivity on at least one of the four texts they translated with TransType. If they were able to do so, it was largely owing to the performance of certain of the prediction engines. In particular, the participants at Celer were able to achieve impressive productivity gains using ITI's English-to-Spanish prediction engine. One translator at Celer more than doubled his/her word-per-minute translation rate using the ITI engine on one of the texts; on another text, the second Celer translator logged the highest productivity of all the participants on the trial, again using ITI's English-to-Spanish prediction engine.

However, when we examine more closely the manner in which the some of the participants actually used the completions proposed by TT2, there is somewhat less cause for jubilation. It seems quite clear that in certain sessions the translators opted for a strategy that is closer in spirit to classic MT post-editing than it is to interactive MT. Instead of progressively guiding the system via prompts toward the translation that they had in mind, the users would often accept the first sentence-length prediction in its entirety, and then edit to ensure its correctness. That the participants were able to increase their productivity by post-editing in this manner certainly speaks well for the translation engines involved. However, our fundamental goal in the TT2 project is to explore the viability of *interactive* machine translation, and this strategy which certain participants adopted – and which is confirmed, incidentally, by replaying the sessions in TT-Player – cannot really be viewed as true IMT.

Still, in our research project as elsewhere, the customer is always right. If the participants at Celer and at Gamma did not make more extensive use of the system's interactive features, it can only be because they felt it was not useful or productive (or perhaps too demanding) for them to do so. Thus, the challenge for the engine developers in the remainder of the TT2 project is to enhance the system's interactive features so that the users will freely choose to exploit them to greater advantage.

In addition to the translations they produced, the participants also provided us with a number of insightful comments about their experience in working with TT2. One user told us, for example, that five alternate completions may be too many, particularly when the differences between them are minimal. The participants also pointed out a number of irritants in the GUI, e.g. the fact that the text does not automatically scroll up when the user reaches the bottom-most segment on the screen; or the occasional incorrect handling of capitalization and spacing around punctuation marks. Although none of these are major, they are a source of frustration for the users and do cause them to lose production time. Finally, the trial appears to validate the decision to base a large part of our usability evaluations on the automatic analysis of the trace files generated in each translation session. Not only is TT-Player able to produce a detailed statistical analysis of each session; it also allows us to verify certain hypotheses by replaying the session, as though we were actually present and looking over the translator's shoulder.

6 Conclusions

It remains to be seen whether fully interactive, target-text mediated MT like that offered by TransType will prove to be a productive and a desirable option for professional translators who are called on to produce high-quality texts. The TT2 project still has more than a year to run and there are many improvements we plan to implement and many avenues that we have yet to explore. One thing is already certain, however, and that concerns the essential role that end-users can play in orienting an applied research project like this one. The translators at Gamma and at Celer have already made important contributions to TT2, both in preparing the system's functional specifications and in helping to design its graphical user interface. And through their participation in the remaining evaluation rounds, it is they who will have the last word in deciding whether this novel and intriguing approach to interactive MT is worth pursuing.

References

Blanchon, H., and Boitet, C.: Deux premières étapes vers les documents auto-explicatifs. In : Actes de TALN 2004, Fès, Morocco (2004) pp. 61-70

Foster, G., Langlais, P., Lapalme, G.: User-Friendly Text Prediction for Translators. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia (2002) pp. 148-155

Hutchins, W. John: Machine Translation: Past, Present, Future. Ellis Horwood Limited, Chichester, United Kingdom (1986)

King, M., Popescu-Belis, A., Hovy, E.: FEMTI: creating and using a framework for MT evaluation. In: Proceedings of MT Summit IX, New Orleans (2003) pp. 224-231

Langlais, P., Lapalme, G., Loranger, M.: TRANSTYPE: Development–Evaluation Cycles to Boost Translator's Productivity. Machine Translation 17 (2002) pp. 77-98

Nyberg, E., Mitamura, T., Carbonell, J.: The KANT Machine Translation System: From R&D to Initial Deployment. LISA Workshop on Integrating Advanced Translation Technology, Washington D.C., (1997)

Wehrli, E.: Vers un système de traduction interactif. In Bouillon, P., Clas, A. (eds.): La Traductique. Les presses de l'Université de Montréal, AUPELF/UREF (1993) pp. 423-432