TransType2: The Last Word

Elliott Macklovitch
Laboratoire RALI, Université de Montréal
Montreal, Canada
macklovi@iro.umontreal.ca

Abstract

This paper presents the results of the usability evaluations that were conducted within *TransType2*, an international R&D project the goal of which was to develop a novel approach to interactive machine translation. We briefly sketch the *TransType* system and then describe the methodology that we elaborated for the five rounds of user trials that were held on the premises of two translation agencies over the last eighteen months of the project. We provide the productivity results posted by the six translators who tested the system and we also discuss some of the non-quantitative factors which influenced the users' reaction to *TransType*.

1. Introduction

TransType2 (TT2) was an international research project, funded under the EU's Fifth Framework Programme, that ran from March 2002 until February 2005. Its participants included three university-based research labs (RWTH in Germany, ITI in Spain, RALI in Canada), an industrial research partner (XRCE in France), an administrative coordinator (Atos Origin in Spain) and two translation service bureaus (Société Gamma in Canada and Celer Soluciones in Spain). The goal of the TT2 project was to develop a new kind of **interactive** machine translation (IMT) system which would help professional translators produce high-quality translations in a cost effective manner. For a detailed description of *TransType's* novel approach to IMT, see (Foster et al. 1997) and (Foster et al. 2002). Suffice it to say here that the focus of the interaction in *TransType* is squarely on the target text, contrary to classic IMT systems, where the user is called upon to help the system disambiguate the source text. Our system observes the user as she types her translation of a source segment and, exploiting an embedded statistical MT engine, it attempts to extend that target translation by proposing one or more completions which are compatible with the prefix the user has entered. Should the user agree with the system's prediction, she can easily incorporate it into her translation by hitting an acceptance key; otherwise, she can edit the proposed string or ignore it entirely by simply continuing to type her target text. However, each new character the user enters provides the system with additional information, which it uses to recalculate and propose a new prediction, all in real time. (See Figure 1 at the end of this paper for a snapshot of a *TransType* session.)

This target-text mediated IMT, as (Foster et al. 1997) called it, certainly is an intriguing idea – but will it work? It should, at least in theory, because each proposed completion the user accepts reduces the number of keystrokes she needs to type in order to produce her desired target translation. On the other hand, the user has to evaluate the system's proposals, i.e. decide on whether to accept them in whole or in part, and this too takes time and effort. Whence the need for a bona fide user trial. In fact, a full-fledged usability evaluation was a major project component in TT2. As its name suggests, this evaluation involved placing the CAT tool we were developing in the hands of professional translators – represented in our consortium by the two translation firms, Celer Soluciones in Madrid and Société Gamma in Ottawa – and having them assess its usability. Contrary to the internal technical evaluations which also figured prominently in TT2 and which employed automatic metrics like BLEU and NIST to assess the performance of successive versions of the translation engines, the goal of these usability evaluations was to gauge the actual impact of the full prototype, including its graphic user interface, on the productivity of working translators, as well as the ease or difficulty with which they adapted to the system. End-users were indispensable to this usability evaluation; for in the final analysis, they are the only ones who can determine whether or not this kind of IMT is really of benefit to professional translators. They have the last word.

Our paper is organized as follows: We will begin by presenting the methodology that we elaborated for the TT2 usability evaluations, describing the protocol that was adopted for the *in situ* trails that were conducted on the premises of the two translation agencies. After providing the productivity results that were registered by the six translators who participated in these extended trials, we will qualify these results somewhat in Section 3 by describing various problems that we encountered in applying our evaluation methodology and some of the lessons we learned in trying to correct them. Section 4 will focus on the evaluation's non-quantitative results; here, we will summarize the users' principal comments and reactions to working with *TransType*. And finally, in Section 5, we will try to provide a synthesis of the overall results of the TT2 user trials and consider the future prospects for this novel approach to IMT.

2. Evaluation Methodology

Although the NLP literature is replete with methodologies for evaluating MT systems, the great majority of these have been designed for fully automatic systems and hence are not entirely appropriate for an interactive system like *TransType*. Many of these methodologies have been inventoried and categorized within the ISLE project, the results of which are now available on-line.[1] Here is the

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

definition we find there of interactive MT: "Interactive MT systems require user guidance at points when the system reaches an impasse during processing." And here is one of the metrics which ISLE proposes for evaluating IMT systems: "measure the amount of time it takes to perform interactive translation on test corpus." This, in essence, was the approach we employed in the quarterly TT2 user trials: we turned the system over to users and allowed them to work with it *in situ*, under real operating conditions; and we carefully clocked the time it took them to complete their translations both and without the benefit of *TransType's* proposed completions.

2.1. *In situ* user trials

Over the last 18 months of the TT2 project, five rounds of user trials were organized on the premises of Celer Soluciones in Madrid and Société Gamma in Ottawa. At each site, two senior translators were selected to participate in the trials; a third was later added for the final evaluation rounds, ER4 and ER5. The first two evaluation rounds were largely preparatory in nature. A *Windows*-based version of the system was installed on each participant's PC and on-site training was provided in order to ensure that the translators were familiar with all of the system's novel features. By allowing the participants to work in their own offices, we wanted them to feel comfortable with the operating environment and also to have access to all the resources that they normally consulted during translation, e.g. dictionaries, glossaries, on-line term banks, etc.

Once the final evaluation protocol was established, each round of trials lasted about two weeks, during which the participants were asked to spend half their working day translating sections of 2000-2500 words using *TransType*. (Half-day working sessions were proposed because initially we were somewhat uncertain about how well *TransType* would perform and how the users would take to this new tool, and we were worried that it might be counter-productive to force the translators to struggle with the system for protracted periods of time.) To gauge the participants' productivity, we simply clocked the time it took them to translate a given source text and then divided that time by the number of words in the text in order to produce a words-per-hour or words-per-minute figure. This is the standard way of measuring translator productivity in the industry, e.g. 350 words per hour.

The corpus that was employed for these *in situ* trials was drawn from a collection of user guides for various Xerox printer and copying machines provided by another consortium member, Xerox Research Center Europe. One of the advantages of this corpus was that all the manuals were available in English as well as in the project's three target languages: Spanish, German and French. From this collection of manuals, a 1 million word corpus was selected, to be used by the three research labs to train and tune their prediction engines; another 40 thousand words were withheld for testing purposes in the user trials. It was decided, moreover, that the participants at the two trial sites would translate the same texts: those at Celer would use the system to translate them into Spanish, while their colleagues at Gamma would translate them into French. XRCE also made available to us a multilingual terminology glossary containing about 750 entries, as well as the English source manuals in their original FrameMaker format.[2] Both of these resources were included in the package that was installed on each participating translator's machine from ER3 onward.

2.2. Trace file analysis

One minor difficulty with the method proposed to measure productivity was how to obtain an accurate and reliable measurement of the time spent on each text. We could, of course, have asked the participating translators to time themselves, but we feared the results would not be entirely reliable; moreover, we didn't want to distract the them from their principal task, which was to translate these difficult texts with the help of this new tool. In the end, we decided to add a trace file to the TT2 GUI, which would record every interaction between the user and the system, and associate with each a precise time stamp. These figures would tell us exactly when the user began to translate a given file and when she completed it. In addition, we added a feature to the GUI which would suspend the system clock automatically after *x* minutes of inactivity on the part of the user. That way, if the participant forgot to manually stop the clock when the phone rang or she went out for a coffee, we could still be confident that the recorded times were relatively accurate.

In addition to translator productivity, we were also interested in examining how the participants actually made use of the system's many modifiable options. And here too the introduction of a detailed trace file proved to be invaluable. In Figure 1, the narrow pane on the right (which the translators would not normally see) shows an extract of a typical trace file. Given a trace of this detail, it is possible to extract a broad range of measurements and statistical indicators from the raw data. Of course, no one would want to do this by hand, since a trace file for a three hour session could easily contain tens of thousands of lines. In order to facilitate the analysis of this data, the RALI developed a utility program called *TT-Player* which takes such a trace file as input and outputs a statistical summary of the session, highlighting whatever statistics we consider important. Some of these are shown on the right side of the lower panel in Figure 1, but the full statistical report automatically produced by *TT-Player* tells us a great deal more, e.g. how many completions the system proposed in a given session and what percentage of these the user accepted; how many of the proposed completions were accepted in their entirety and many were accepted in part; the average number of words in the accepted completions; the average time required to accept a completion; how many completions were accepted via the keyboard and how many using the mouse; how many characters in the final target text were actually typed by the user and how many derived from system completions; etc. *TT-Player* can also function like a VCR, allowing us to replay a translation session and observe exactly how the translator uses the system's predictions to construct her target text. For more on *TT-Player*, see (Macklovitch et al. 2005).

Actually, more than the system's impact on translator productivity, what really interested us was to determine whether the participating translators would be able to increase their productivity with the help of the

[2] Input to *TransType* must be in plain text format. The Framemaker manuals allowed the participants to view graphics and images that could help with the translation.

TransType's completions. Although both translation firms already had average productivity figures for each participant, we wanted to ensure that all possible variables in our trials were kept constant, so that the numbers we obtained would be an accurate reflection of the contribution of *TransType's* predictions and nothing else. To that end, each user trial included what we called a 'dry-run' session, during which the participants were asked to translate a chapter of the test corpus on their own, i.e. using *TransType's* editor but without the benefit of the system's completions. This dry-run session provided us with baseline productivity figures against which we could then compare the participants' productivity on the same type of technical manuals but translated with the help of the system's proposed completions. Table 1 on the last page of this paper gives the six participants' average productivity figures on the final three evaluation rounds.

3. Discussion of Productivity Results

ER3 marked the first time that the participating translators at Société Gamma and at Celer Soluciones had the opportunity to actually work with the TT2 system in a mode that approximated their real working conditions. However, this evaluation round only lasted five half-days, including one for the dry-run. Furthermore, over the remaining four half-days, we asked each participant to test two quite different system configurations: one in which TT2 generated shorter, multiple completions and the other in which it generated a single, full-sentence completion. (It turned out that the users expressed a clear preference for the latter configuration, saying that having to read through and evaluate multiple predictions caused them to lose an undue amount of time.) Hence, one shouldn't attribute too much significance to the ER3 productivity figures in the second row of Table 1. We found it sufficiently encouraging that even with this change in configurations, three of the four participants actually managed to surpass their dry-run productivity rate on at least one of the texts they translated with the system's predictions.

The fourth round of user trials (ER4) took place in late July – early August 2004 and represented a major scaling up of the evaluation process in several respects. For one thing, we wanted to produce more translations with TT2, and so a third participant was added at each test site. All six participants were asked to translate eight new sections of the Xerox test corpus, as opposed to the three sections that were translated at Gamma in ER3 and the four that were translated at Celer. The timetable that was proposed for ER4 therefore called for ten consecutive half-day sessions (instead of the five that had been suggested in ER3), with the first session devoted to refresher training and the second to the dry-run. Including the dry-run text, the test corpus for ER4 totalled 15,420 words. As for the productivity results obtained on ER4, our initial analysis of the logfiles showed them to be extremely impressive: five of the six participants exceeded their dry-run productivity on seven of the eight texts they translated using TT2's completions, and three of these did so on all eight texts. This, despite the fact that the dry-run rates on ER4 were higher for three of the four translators who had participated in ER3. In short, TT2's predictions seemed to allow almost all our participants to translate these chapters of the Xerox printer manual substantially faster than they could on their own. What's more, an independent revision of the translations produced with the help of the system's predictions showed that they contained no more errors than the translations of the dry-run; in fact, all were considered to be of deliverable quality. This too was an important finding, since it confirmed that the impressive gains in productivity were not obtained at the expense of translation quality.

After the fourth round of user trials had been completed, however, we discovered, somewhat to our consternation, that there was a high degree of full-sentence overlap between the eight portions of the test corpus for that round and the corpus that had been used to train the prediction engines.[3] This overlap was not in the actual chapters from which the test and the training corpora were drawn; in other words, no oversight or error had been committed in selecting the test corpus. Rather, it was the result of the highly repetitive nature of these Xerox user guides: within different chapters of the same manual, or even across different manuals for similar types of equipment, identical commands and the same sentences reoccur verbatim over and over again. Nevertheless, we decided to reanalyse the ER4 results, scrupulously separating the repeated sentences from the singletons and extracting from the traces files the corresponding data for each. What we found, not surprisingly, was that translator productivity on the repeaters was substantially higher than on the singletons. Although the embedded SMT engines did not strictly speaking incorporate a translation memory's string matching capability, the system's initial predictions on most of the repeaters in the ER4 test corpus were generally of excellent quality; the translator could often accept them as is and then, if necessary, make a few minor post-editing corrections. When we calculated the average productivity of the six participants on the sub-texts composed only of singleton sentences and compared it to their dry-run productivity, we still obtained a 20% productivity gain across the board. This is less than the 32% average gain that we had initially reported, but it is nevertheless far from negligible

Be that as it may, in the fifth round of user trials (ER5), we went to great lengths to ensure that *none* of the sentences in the files retained for the ER5 test corpus appeared verbatim in the training corpus.[4] And another important change was introduced in the evaluation protocol for ER5: we added a second dry-run session, scheduled near the end of the ten-day trial period, during which the participants were asked to translate another section of the test corpus with the system's prediction engine turned off. This was to counter the argument that the dry-run productivity figures that were used as a baseline in ER4 may have been unfairly low, since the one dry-run session in that round had been scheduled on the first day and the translators' performance seemed to gradually improve over the ten-day trial period.

The participants' results for ER5 appear in the bottom portion of Table 1. Perhaps the first thing to notice is that, when we compare the translators' average productivity gains on the eight texts they translated using the system's predictions with their productivity on the first dry-run

[3] By one account, 41% of the sentences in the ER4 test corpus appeared verbatim at least one time in the training corpora.

[4] In retrospect, this concern seems somewhat exaggerated, since repetitions are a salient characteristic of these types of manuals.

(DR1), the increases are more modest than those that were registered on ER4. And in absolute terms too, the participants' average word/hour rates are not as impressive as those posted on ER4; for three of the six participants, these actually drop on ER5 and collectively, their combined productivity rate is significantly lower on the latter round than on the former (995.7 w/h on ER5 vs. 1111.5 w/h on ER4.) Recall, however, that the figures for ER4 given in Table 1 include the high percentage of full-sentence repetitions; it would perhaps be fairer to use as a comparison with ER5 the productivity increases that the participants obtained on the sub-corpus of ER4 which included no repeaters. When we calculate the average productivity of the six participants on the ER4 sub-corpus that is composed solely of singleton sentences, we obtain a figure of 16.75 words per minute. And if we do the same for the eight texts translated by the six participants in ER5, the figure we obtain is 16.65 words per minute. So this comparison – which abstracts away from an important difference in the make-up of the two test corpora – suggests that the participants' productivity over the two evaluation rounds was actually a fairly constant.

The real surprise in the ER5 results came in the figures the participants recorded on the second dry-run (DR2). On this text, the participants' collective average productivity was 1346 w/h, much higher than their average on DR1 (928 w/h) or on the eight texts they translated with the help of the system's predictions (996 w/h). Indeed, when we combine each participant's individual productivity on DR1 and DR2 and compare the result with her average productivity on the eight test texts, it turns out that none of the five participants[5] showed any gain in productivity on ER5 – contrary to the encouraging gains they had posted on ER4. All this seemed to confirm our fear that we had failed to take into account the learning curve effect, whereby the participants' performance gradually improved over the course of an evaluation round. On this second dry-run, scheduled near the end of the 10-day trial period, three of the five participants logged their single best productivity figure for the whole evaluation, and for the other two translators it was their second or third best. Upon closer examination, however, we again found that the situation was more complex than it appeared at first blush. In segmenting the test corpus into 2000-word chunks, we had blithely assumed that all the resulting portions would be of more or less equal difficulty, seeing that all were drawn from a similar set of Xerox manuals. But as it turned out, the text that we inadvertently selected for DR2 was much easier to translate than all the others in that round. We later measured the average length of the sentences in this text and discovered they were shorter than those of any of the other test files in ER5.[6] Moreover, a full 27% of the sentences in the text were internally repeated at least once, and some up to ten times. All things being equal, a text which contains more internal repetitions should be easier for a translator than one which contains no or few repeated sentences. Further evidence that this DR2 text was not representative of the general level of difficulty of the test corpus came from the participants' productivity figures on the final text they translated in ER5, following the second dry-run: the productivity of four of the five participants declined significantly on this final text, contrary to what one would expect if the learning curve effect was the sole or determining factor.

4. Non-quantitative Results

One way of viewing the problem we encountered on the second dry-run text in ER5 is this: Had we selected some other text for DR2, one that was more representative of the overall level of difficulty of the test corpus, then we probably would have arrived at a different conclusion regarding the system's impact on productivity in this last trial round. In terms of our general evaluation methodology, there is certainly an important lesson to be learned here, which is that it is critical to assess and control the difficulty of the texts that are used in the test sets. Be that as it may, there were other evaluation rounds where the participants' results were non-problematic and entirely unequivocal, e.g. those of ER4 and even ER5 using DR1 as a baseline comparison. In these cases, the figures clearly show that the use of *TransType's* predictions did allow the participants to increase the rate at which they produced their translations. The gains in productivity may not have been spectacular, but they are certainly significant.

As we mentioned in the Introduction, productivity was not the only parameter we were interested in evaluating in these user trials. We also wanted to gauge the ease or difficulty with which the participants adapted to the system. More generally, we were interested in any and all comments that they might have to make on any aspect of the system's use. In order to encourage the participants to share their spontaneous reactions with the developers, we added a pop-up notepad to the system's GUI, which the users could easily call up via a keyboard shortcut (thereby halting the system clock). Entries were automatically time-stamped and identified with the user's name and later compiled into Comments files that were scrutinized at the end of each evaluation round. The suggestions and complaints in these Comments files proved very helpful in allowing the developers to make corrections and improvements to the prototype in time for the following evaluation round.

This is clearly not the place to present all the remarks the users inscribed in the system's notepad. There were, however, certain recurrent comments and complaints, two of which we will mention here because they had a definite impact upon the users' attitude of the system. The first had to do with the problem of repetitions, which, as we have seen, was a salient characteristic of the Xerox manuals from which our test corpus was drawn. In particular, internal repetitions were especially frequent in the DR2 text.[7] When the system's initial prediction on these sentences was not to the translators' liking, they would modify it a first time; and later, when that same sentence re-occurred within the file, they found they had to make the same corrections over again. This was something they did not at all appreciate, as they made very explicit in their comments. In an interactive tool like *TransType*, this is a

[5] G-TR6 did not translate either of the two dry-run texts in ER5, making it impossible to determine if he showed any increase in productivity using the system's predictions on this round.

[6] There is a well-known rule of thumb in MT which correlates the difficulty of a sentence with its length.

[7] As opposed to external repetitions, involving sentences in the test corpus that are repeated in the training corpus. These, the users were not overtly aware of.

clear lacuna: the system needs to incorporate a simple string matching and repetitions processing capability like that found in most commercial translation memory systems, particularly since translators have now become accustomed to working with translation memories.

A second complaint which the users repeatedly voiced in their comments appears, at first glance, to be quite similar to the problem of repetitions processing. The participants again complained about the fact that system failed to take account of their corrections, except in this case the corrections in question did not occur within repeated sentences. The distinction is subtle but important: If a corrected sentence re-occurs later in the text, it should not be difficult in principle to recall its corrected translation; this is just what simple repetitions processing allows for. However, suppose that the translator makes a correction in a sentence which does not re-occur verbatim, but that later in the text just the one or two words she corrected do reappear. A simple translation memory won't solve the problem here. Moreover, there is a good likelihood that *TransType* will reproduce the problem which the translator initially corrected every time it reoccurs, since the system's underlying language and translation models remain unchanged during a working session. This too, the participants found particularly frustrating. "Why can't the system learn from my corrections?" they asked over and over again. The answer is simple enough: The prediction engines embedded in *TransType* do not incorporate on-line, adaptive learning capabilities; the parameters of their language and translation models are acquired prior to the interactive sessions, via computationally intensive training cycles. Unlike the solution for full-sentence repetitions, however, implementing the changes that would correct this problem, is anything but simple. Indeed, the problem of how to endow a SMT engine with adaptive learning capabilities, so that it could learn from the user's corrections and modify its underlying models appropriately, would necessitate a major research effort.

5. Prospects and Conclusion

The user trials that we have described above were not the only type of evaluation conducted within TT2. As mentioned above, the project also included internal technical evaluations conducted by the three university labs developing the system's prediction engines, where the goal was to assess the relative performance of successive versions of the prediction engines, i.e. to ensure that the quality of the predictions was improving from one round to the next. Over the course of the 3-year project, there were indeed noticeable improvements in prediction engine performance. More generally, TT2 did give rise to significant advances in the field of SMT; among a host of publications, see, e.g. (Bender et al. 2005) and (Civera et al. 2004).

These improvements in translation engine performance were certainly an important factor in determining how the users reacted to the system. As one would expect, the better the quality of the predictions generated (as on ER4, for example), the happier the participants were in using the system, because the more work it did for them. However, engine performance was not the only factor affecting our users' attitude to *TransType*. As we have seen, even on those trials where the prediction engines were performing remarkably well, there were other usability factors that came into play, e.g. the system's inability to retain previous revisions made by the translator, which could sour the user's attitude to the system. As developers of CAT technology, our role is not to downplay these kinds of user reactions, but rather to learn from them. And what the translators told us in no uncertain terms during these usability trials was that unless some way could be found to avoid their having to make the same corrections to *TransType's* output over and over again, they were not interested in employing the system. And this, despite the fact that *TransType* allowed them to obtain significant gains in productivity – on the order, say, of 15-20% on non-repetitive texts.[8]

The TT2 project ended in February 2005 in Europe and six months later in Canada. Since that time, no further work has been done on the system; in particular, no attempt has been made to respond to our trial participants' principal complaint regarding the system's inability to retain and exploit the revisions they made to its output in order to improve the quality of subsequent predictions. The question which naturally arises, then, is whether it would be possible to correct this failing without having to engage in a major new round of research. My inclination is to answer in the affirmative, at least for complete sentences, for which it would not be difficult to add a TM-like repetitions processor to *TransType*. Would that be sufficient to overcome our participants' reticence to use the system on a daily basis? Without an additional round of user trials, it is hard to say; because, as we stated at the outset, it is only the intended users of the system who can determine whether this novel approach to IMT is of benefit to them or not. They always have the last word.

6. Acknowledgements

I wish to express my gratitude to all the researchers who contributed to the TT2 project, and particularly to the translators who participated in the user trials.

7. References

Bender, O., Hasan, S., Vilar D., Zens, R., Ney, R. (2005). Comparison of Generation Strategies for Interactive Machine Translation. In Proceedings of 10th Annual Conference of the European Association for Machine Translation (pp. 33--40). Budapest, Hungary.

Civera, J., Vilar, J.M., Cubel, E., Lagarda, A.L.,. Barrachina, S., Vidal, E., Casacuberta, F., Picó, D., González, J. (2004). From machine translation to computer assisted translation using finite-state models. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (pp. 349--356). Barcelona, Spain.

Foster, G.., Isabelle, P., Plamondon. P. (1997). Target-text mediated interactive machine translation. Machine Translation, 12(1-2), 175--194.

[8] These numerical estimates need to be taken with a grain of salt, for this remains a relatively small-scale trial. Over the course of the project, only six translators actually worked with *TransType*, none of whom used it to translate more than 40 thousand words.

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

Macklovitch, E., Nguyen, N.T., Lapalme, G. (2005). Tracing Translations in the Making. In Proceedings of MT Summit X (pp. 323--330). Phuket, Thailand.

	C-TR1	C-TR2	C-TR3	G-TR4	G-TR5	G-TR6
ER3: dry-run (words/hour)		984	432	864	786	
ER3: average on 3-4 texts (w/h)		918	774	882	576	
ER4: dry-run (w/h)	781	1030	772	518	1081	825
ER4: average on 8 texts (w/h)	1017	1410	725	707	1531	1279
% increase in productivity	+30.22	+36.89	-6.08	+36.48	+41.62	+55.03
ER5: dry-run1 (w/h)	924	858	654	864	1338	
ER5: average on 8 texts (w/h)	1056	1104	736	1104	1062	912
% increase in productivity	+14.29	+28.67	+12.54	+27.78	-20.63	
ER5: dry-run2 (w/h)	1602	1416	816	1548	1350	
% increase in productivity	-34.08	-22.03	-9.80	-28.68	-21.33	
ER5: average on 2 dry-runs (w/h)	1290	1137	735	1206	1344	
% increase in productivity	-18.1	-2.9	0.0	-8.4	-20.9	

Table 1: **Average Productivity Results on ER3, ER4, ER5**

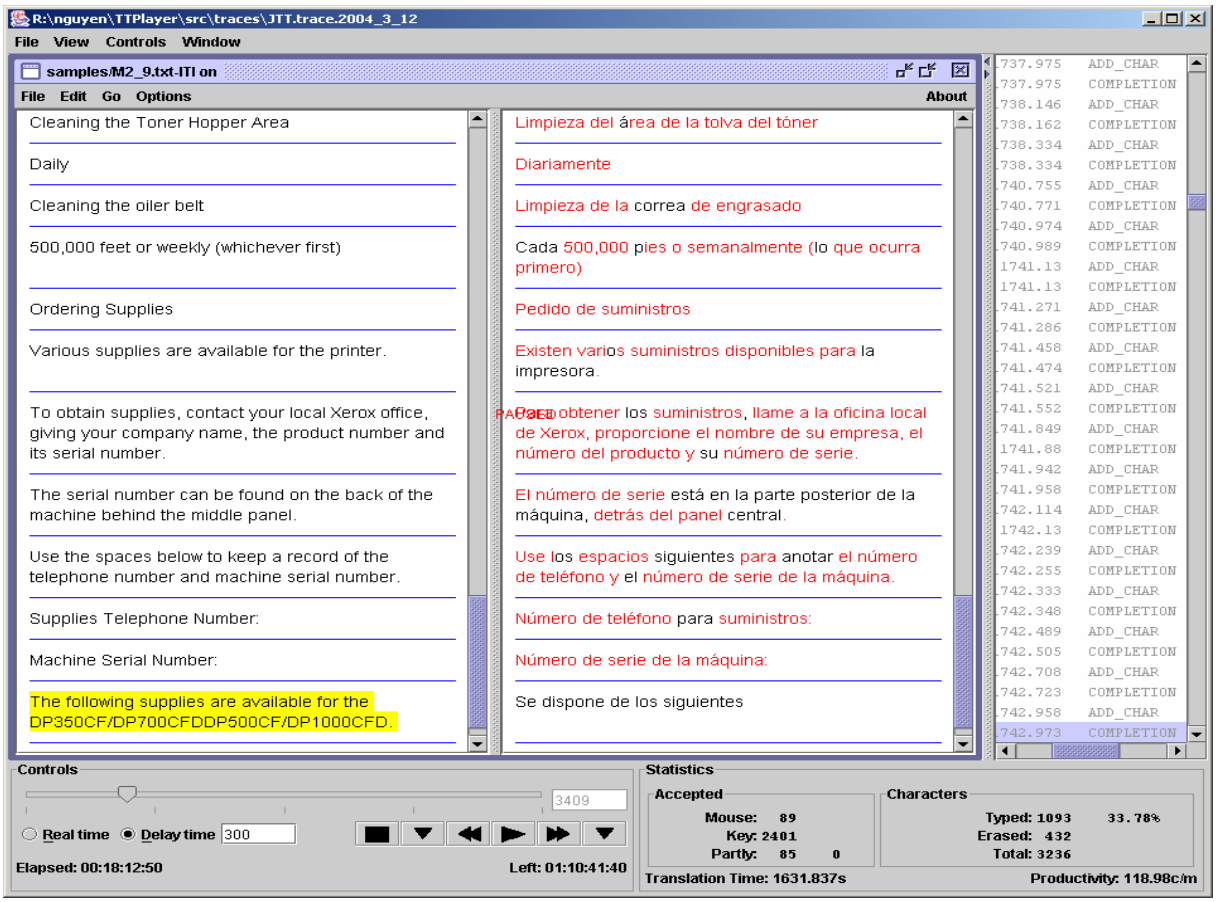

Figure 1: **Snapshot of TT-Player** in replay mode (English to Spanish translation)
The light characters derive from the *TransType's* predictions; the dark ones were entered by the translator