An Experiment in MT Post-Editing by a Class of Intermediate/Advanced French Majors.[1]

Michael Kliffer

McMaster University
Hamilton, Canada
kliffer@mcmaster.ca

Abstract. This paper describes an experiment whose goal was to introduce MT via post-editing to a third-year university class in French-to-English translation. Even though most of the students in our French programs do not go on to become professional translators, it seemed a worthy objective to give the[2]m at least a taste of the challenges and opportunites which MT offers today, given how widespread it has become in the commercial and technical sector. It would also be valuable to know to what extent a brief hands-on encounter with MT would help them better understand the nature of translation. After providing background on the course content, level of students and type of documents chosen for the experiment, I briefly look at the question of how to "teach" MT and give reasons for the focus on post-editing. I then outline the experiment itself, including the process of choosing an MT program, the type of texts, the error analysis adopted for evaluating both human and MT output of those texts, a comparison of human and MT results and finally student reaction to the experiment. I conclude with a glance at the next phase of this project, a parallel application in English-to-French translation, where, unlike the first phase, the target language is not the students' native tongue.

1. Background: course, level of students, type of documents

The course chosen for the experiment is a level 3 French-to-English course which is one of four translation courses intended as a supplement to language courses rather than part of a translator-training program. This distinction is important to note for the experiment because the course's general orientation means that the focus is on mainly journalistic, topical texts instead of the specialized subjects expected in a course for training professionals. The course continues the presentation of basic translation skills, such as proper dictionary usage, context-based inferencing strategies, and English-French contrasts in lexicon, style, and advanced syntax. As the only one of our translation courses working exclusively from French to English, the native language of most of our students, it places more emphasis on precise interpretation of the source text.

The students have typically had several years of French in secondary school, so that by third year in our program their French is quite advanced. Our courses immerse them in French, whether the content is literature, linguistics, culture/civilization, or language-based. Yet, the course where the experiment was carried out is the first time most are exposed extensively to journalistic texts and in spite of their extensive study of literature, they find journalistic translation quite a challenge, thanks largely to the abundant cryptic references, the cultural subtext, and the often opaque figurative language. Of course, this is not to downplay target text

[1] [1] This paper is a revision of Kliffer 2001 which describes the experiment in French. I have added a discussion of post-editing, focusing on the differences between its use in commercial/technical translation and its exploitation in the classroom.

factors, such as the audience's existing beliefs and encyclopedic knowledge, but several years of teaching this course have confirmed that the biggest challenge for all but the best students is to arrive at a correct understanding of the source text.

2. "Teaching" MT: why focus on post-editing ?

MT is one of five ancillary topics on translation theory that the students read about as a complement to their regular "hands-on" work. Before the students attempt post-editing, they read Hutchins 1992 and Sampson 1987 order to learn about the major prototypes of MT, its chequered history, controversial claims and the reasons why many commercial and governmental organizations now consider PE indispensable for producing publishable-quality output.

An experienced translator may well question the appropriateness of simulating PE in a course on general translation. As Melby & Warner 1995 insist, successful MT today is confined to texts which utilize sub-languages restricted to a particular, usually technical, domain. This is the diametric opposite of journalistic texts, aimed at a general public and free of any such constraints. Moreover, PE specialists stress efficiency, which entails that the revisor should not devote too much time to stylistic niceties and should avoid the temptation to rewrite throughly the output text, caveats going against the methodology of our translation courses, which encourage the student to aim for semantic and stylistic precision via multiple rewritings.

Nevertheless, there are good reasons for having students perform PE in order to learn about MT. First, it is arguably the component of MT that ties in best with translation desiderata already emphasized within the course, notably concern for semantic and functional accuracy vis-à-vis the source text. The other components would entail either practices that contravene the course methodology (e.g. pre-editing, which suits technical writing but in most cases not journalistic texts) or abilities beyond its scope (e.g. writing MT code). On the other hand, revising MT output is feasible: if students have access to the output and the source text, they can apply their existing knowledge of both languages to make corrections, without any special technical training. PE thus makes student aware of MT's capabilities and limitations, as well as providing additional translation practice through revising.

3. The experiment

3.1. Choice of MT software

In view of the intermediate/advanced level of the course, it seemed pointless to use a gisting program like *Babelfish*, which would produce too many low-level errors that a human would be unlikely to make. The other extreme, i.e. a high quality program requiring little post-editing, even if it existed, would not have had much pedagogical value either. Intermediate-level software, luckily, was readily available.

My Internet searches and perusal of both commercial and academic software evaluations led to purchase of a single package, *Power Translator Pro*, henceforth PTP. I also tested demonstration versions of Softissimo (downloaded) and Systran Professional (on-line translation). None of these programs could be considered cutting-edge MT, all three being either the direct or transfer type.

Evaluations of PTP were nonetheless impressive. From a review in *PC Computing* to a master's thesis (Justice 1998), critics sang its praises thanks to the quality of its output, the time saved (35-40%), its flexibility and the number of languages offered in a single package (5, in addition to English). When I tried out PTP, the results confirmed how far short of expectations the program fell. Despite its sentence-by-sentence approach, PTP's output requires intensive post-editing largely because its syntactic treatment of lexical items, especially verbs, is very uneven. In its performance, PTP nevertheless proved to be on a roughly equal footing with *Softissimo* and *Systran*, which were far more costly.

3.2. Error Analysis and Comparison of MT and Student Output

3.2.1. Placement Test

Although the first stage of the project deals with translation into English, I started by evaluating the 3 MT programs via a French grammar placement test. Our department used this test for 7 years in order to determine the appropriate course for students entering 1st year; this provided us with a precise idea of the average level of students coming out of high school. The test consists of 50 multiple choice questions, incorporated into a continuous monologue. They cover grammar points which the student is supposed to have mastered in secondary school, such as all the verb tenses of the spoken language, the main difficulties of preposition usage, and the principal morphological irregularities.

In order to adapt the test to the MT programs, I translated the monologue into English so as to focus on the same points as in the original format. The objective of this translation was to see if the programs could handle the same grammar problems as the students had faced. This would allow two evaluations: a comparison between the programs' and the students' performance, and a comparison among the programs themselves.

Here are the results: Placement Test (Score/50): Student median 28; PTP 29; Systran 30; Softissimo 32

The MT scores reflect errors only on the same points that the students were tested for.

To explain these rather low scores, we need to consider the following factors. Most of the students had spent at least 2 months without any contact with French and took the test, entirely written, without any special preparation. The MT programs too had "taken" the test under less than ideal conditions. Even though PTP and Systran offer an interactive mode which would have allowed a human translator to make lexical and grammatical choices, the programs handled the input without any human intervention. In a real situation, the translator could have benefitted from other resources like online dictionaries and translation memory and, if the text had been of a commercial or technical nature, would have likely done some pre-editing.

With regard to errors, the main difference between the software and the students lay in the morphology/syntax distinction. While the students' errors stemmed about equally from both components, the programs' errors were almost entirely non-morphological, which reflects mastery of the closed system of French inflexion. MT weaknesses appeared especially in discontinous syntactic dependencies, e.g. mood choice of an embedded verb, which is determined by the main clause verb.

The results of our placement test show little divergence between human and machine: the range between the students' average of 56% and the highest MT mark of 64% is not large. Moreover, the programs obtained similar results to each other, in spite of considerable price differences. Their scores suggested that they would indeed provide output that would lend itself well to post-editing by an intermediate/advanced class.

3.2.2. The texts

Three texts were used:

1. *Le sida devient maladie de pays pauvres* ('AIDS is becoming a third-world disease', Interview with Luc Montagnier, *L'Express*, March 26, 1998)
2. *Quand les immigrants regardent notre petite vie* ('When Immigrants view our day-to-day lives', TV and new Quebecers, *L'Actualité,* June 1, 1997)
3. *À vous de jouer* ('It's your play', The nightmare of buying Christmas presents for today's kids, *Le nouvel Observateur*, December 5, 1996)

Students from a previous year had already translated these texts. I had each one translated by the 3 programs used for the placement test evaluation. A detailed comparison showed that the 3 were roughly equivalent with respect to frequency and nature of errors. I thereafter restricted the project to PTP, with the aim of comparing its mistakes with those of students.

3.2.3. The error analysis

I had originally intended to use the error classification of the Society of Automotive Engineers, whose model has served in numerous evaluations of technical translations. Realizing that its grammatical/lexical categories were too general to adequately capture PTP's problems, I had to work out a 'made-to-order' package of criteria.

Most of the categories, such as *verb-preposition, tense,* and *word order* are transparent, but 3 categories require some explanation. The first, *word choice*, includes polysemy and homonymy. It nearly always involves a co-textual element which should have triggered a different lexical choice. This element is at times a headword with a semantic feature that conflicts with PTP's lexical choice:[2]

(1) La télévision francophone: French-speaking television (–> French-language)

It may also clash with a feature of the subject:

(2) Mais toujours la liste de l'enfant prime: ... But the child's list always excels (–> has priority)

or with the verb:

(3) Ils (les fabriquants de jouets) font monter savamment le stress de Noël: They (the toy manufacturers) learnedly make rise Christmas stress... (–> cleverly)

Since the word choice often depends on a discontinuous element, generally impossible to pin down via morpho-syntactic criteria alone, this is likely the most challenging category for MT.

The next category of error is *anaphor*. This includes classic reference problems like

(4) ...ils préfèrent les Etats-Unis. Dans ce pays...: ...they prefer the United States. In this country...

where English prefers 'that' when the antecedent has already been mentioned. This category also includes any gender or number mismatch with personal pronouns, since we are still dealing with nouns already given in the discourse:

(5) ... le sida est dû à un virus. Sans lui, il n'y aurait pas d'épidémie.: ...the AIDS is due to a virus. Without him there would not be an epidemic.

as well as difficulties in the usage of adverbial pronouns *y* 'there' and *en* 'of/from it, some':

(6) Si vous êtes en Afrique, vous en crevez.: If you are in Africa, you burst some. (–> ...you die from it (AIDS))

The final non-obvious label is the catch-all *mistranslation*, applied to errors for which there is no apparent explanation:

(7) Aux gosses de cet âge qui demandent des jouets...: To the youngsters of that age that ask for the toys... (–> ...ask for toys)

3.2.4. Comparisonwith human translators.

Before discussing the data, I should mention an important point about the student translations. They deal of course with the same three texts, but I analyzed the output of a different student for each text. All 3 cases involved students in the lowest quartile because one of the aims of the project was to bring out typical errors of weak students.

Power Translator Pro

	N	%
word choice	334	30.7
literal	167	15.4
structure	110	10.1
generic	101	9.3
preposition	77	7.1
anaphora	64	5.9
word order	55	5.1
tense	41	3.8
prep-verb	36	3.3
omission	30	2.8
mistranslation	24	2.2
not found	24	2.2
agreement	11	1
prep-adj	9	0.83
number	3	0.28
article	2	0.18

Total errors: 1088

Total words: 4754

Table 1: PTP Errors

Student Translations

	N	%
word choice	130	26
mistranslation	93	19
literal	60	12
omission	40	8.1
spelling	38	7.7
tense	28	5.7
preposition	22	4.4
punctuation	17	3.4
anaphora	17	3.4
generic	14	2.8
structure	14	2.8
agreement	7	1.4
word order	6	1.2
article	4	0.81
redundant	3	0.6
not found	2	0.4
prep-verb	1	0.2

Total errors: 496

Total words: 4598

Table 2: Student Errors

[2] The example translations obviously show more errors than just the one being illustrated.

Tables 1 and 2 compare student and MT errors for 17 categories. The students were in a previous year's class and had translated the articles on their own, with the help only of dictionaries and style manuals. In summary, the MT and human translations share only *word choice* and *literal* as error categories with double-digit percentages. PTP shows high incidence for *structure* and *generic* (i.e. misinterpretation of the generic or specific sense of the definite article), categories which posed few problems for the students, probably because of their anglophone status. The reverse is observed for *mistranslation*, at 19% for students vs. 2.2% for PTP. We see that the lexicon, with its ever-present polysemy, is a challenge for both humans and machine, whereas morphological tasks like agreement are well handled by both.

3.2.5. The post-editing exercise and student evaluations of it

I gave the source text of the Montagnier interview and its PTP translation to 11 students who post-edited it. This group was different from the one which had done the initial translation. I then counted the errors of one strong, one average and one weak student.

Table 3 indicates a marked decrease in errors for the post-edited versions, though the spread among the 3 students group is large: 118, 53 and 12 errors for the weak, average and strong students respectively, yet all showing marked improvement over PTP's score of 374. For weak and average students, *word choice* and *literal* were the most common error categories, just as with the student translations done from scratch.

The last stage of the project comprised a student evaluation of the post-editing exercise. 3 questions were asked. The student was first asked to gauge the usefulness of the exercise, via a 5-point scale. Of the 11 students, 4 chose 'quite useful' (second highest point) and 4, 'somewhat useful' (middle point). The other answers were split between the 2 extremes.

Students were then asked to indicate advantages and drawbacks of the exercise. The pluses focused on the more challenging errors, notably idioms and figures of speech, while the negatives concerned the plethora of "stupid" mistakes, typically over-application of structural principles and reference errors that pragmatic inferencing would allow a human to avoid.

The third question asked for suggestions to improve the exercise. Some samples:

"We should translate the text ourselves before we correct the program's translation."

"You should have let us do the correcting in groups of 3 or 4."

PTP		**Weak St.**	**Average St.**	**Strong St.**				
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
word choice	113	30.2	37	31.3	21	40	3	25
generic	47	12.6	13	11	5	9.4	0	0
literal	44	11.8	18	15.2	10	18.9	3	25
structure	37	9.9	7	5.9	2	3.8	0	0
preposition	34	9	5	4.2	4	7.5	1	8.3
anaphora	26	6.9	3	2.5	0	0	1	8.3
word order	22	5.9	5	4.2	0	0	0	0
tense	19	5	6	5	2	3.8	0	0
prep-verb	12	3.2	0	0	0	0	0	0
spelling	0	0	8	6.7	3	5.7	1	8.3
not found	10	2.7	3	2.5	3	5.7	2	16.6
punctuation	0	0	5	4.2	1	1.9	1	8.3
prep-adj	8	2.1	0	0	0	0	0	0
mistrans.	5	1.3	8	6.7	2	3.8	0	0
agreement	1	0.2	0	0	0	0	0	0
TOTAL	374		118		53		12	

Table 3: Errors in PTP Output vs. Post-editing by Students for Montagnier Text

"It would be a good idea to indicate to us in advance mistakes in the program's translation."

Overall, the weaker students appreciated the exercise more than the stronger ones, some of whom found there were too many "stupid" errors. For the weak ones, evaluating a translation and correcting its mistakes proved to be less stressful than doing a translation entirely by themselves.

4. Conclusion

The experiment was undertaken with full awareness of the current insight that MT is suited primarily to fixed-domain texts, those manifesting what Melby & Warner 1995 call **superficial** ambiguity, where a fixed number of senses are mapped onto a given lexical item. The journalistic texts whose MT output the students post-edited were all in the general, dynamic language domain, where, according to Melby & Warner, **fundamental** ambiguity predominates because of unpredictable meanings arising in novel situations. In a sense, the students' experience with this post-editing task recapitulated the mid-20th century attempt to attain high quality MT with any kind of text. The resulting frustration, seen especially in the stronger students' comments, recalls the reaction of 1970's MT specialists, who eventually realized the fundamental incompatibility between MT capability and the semantically open-ended, indeterminate nature of general, non-technical input. Thus, while the experiment familiarized students with MT and its limitations, it also gave them insight into some fundamental properties of human language.

The next stage in this project is a post-editing exercise where the target language, French, is not the native language of the majority of students. Preliminary results suggest that, as expected with translation into a non-native language, the overall student performance is below that obtained with the texts discussed in this paper, but that awareness of MT's inability to handle figurative and idiomatic language is just as high.

5. References

HUTCHINS, W. John. 1992. General Introduction and Brief History. In *An Introduction to Machine Translation*. W. John Hutchins & Harold Somers London: Academic Press.

KLIFFER, Michael. 2001. Exploitation pédagogique de la traduction automatique. *Distances*. 5.2:83-102.

MELBY, Alan & C. Terry WARNER. 1995. The possibility of language: A discussion of the nature of language, with implications for human and machine translation. Amsterdam: Benjamins.

SAMPSON, G. 1987. „MT: A Nonconformist's View of the State of the Art". In *Machine Translation Today*. M. King (Ed.) Edinburgh: Edinburgh University Press.