Lilt vs. SDL Trados: Productivity Pilot Study

Spence Green
Lilt Inc.
Palo Alto, CA 94309, USA
spence@lilt.com

Abstract

We compare human translation performance in Lilt to SDL Trados, a widely used computer-aided translation tool. Lilt generates suggestions via an adaptive machine translation system, whereas SDL Trados relies primarily on translation memory. Five in-house English–French translators worked with each tool for an hour. Client data for two genres was translated. For user interface data, subjects in Lilt translated 21.9% faster. The top throughput in Lilt was 39.5% higher than the top rate in Trados. This subject also achieved the highest throughput in the experiment: 1,367 source words per hour. For a hotel chain dataset, subjects in Lilt were 13.6% faster on average. Final translation quality is comparable in the two tools.

Introduction

In professional human translation there are two variables of interest: speed and quality. Cost is typically a function of these two variables. The goal of computer-aided translation (CAT) tools is to increase speed while at least maintaining—or in the best case improving—quality. The core feature of all modern CAT tools is translation memory (TM), which is a mechanism for retrieving previously translated segments from a database. TM was first proposed by [1] as a simple way to boost throughput when source text is repetitive. Most TMs

	Condition	Genre
	Morning	
1	Lilt	H
2	Trados	UI
3	Lilt	H
4	Trados	UI
5	Lilt	H
	Afternoon	
1	Trados	UI
2	Lilt	H
3	Trados	UI
4	Lilt	H
5	Trados	UI

Table 1: Productivity experimental design. Randomized assignment of subjects (1–5) to translation conditions and data genres (**UI**: user interface; **H**: hotel chain).

query for both exact and partial matches via an approximate string matching procedure. Other translation aids such as dictionaries, concordances, and, more recently, machine translation (MT) output, augment the TM results.

Lilt is a new CAT tool based on translation productivity research at Stanford University [5]. The interface is designed according to mixed-initiative principles [6] for human/machine systems and is primarily informed by eye-tracking research. Suggestions are provided by an adaptive MT system, which can predict completions for each partial translation typed by the translator. New complete translations are then immediately added to the MT system. In this way, the MT system can personalize its suggestions.

Summary of Productivity Results

This case study reports a pilot productivity experiment with five in-house English–French translators. Five rounds of usability experiments in which subjects were asked to translate text under observation preceded the sixth round, which was a controlled throughput experiment. The baseline condition was SDL Trados, a state-of-the-art CAT tool that the subjects use daily at their employer.

Subjects translated two different genres: user interface (UI) strings from a software product and a hotel chain loyalty brochure. For the UI data, subjects translated an average of 21.9% more source text with Lilt. For the hotel data, subjects translated 13.6% more source text. Across all conditions and genres, the highest throughput was achieved with Lilt: 1,367 source words per hour. Analysis shows that final quality is comparable in the two tools even at this high throughput.

Experimental Design

We conducted a language translation experiment with a 2 (translation conditions) x n (source sentences) mixed design. Translation conditions (SDL Trados and Lilt) and source sentences were the independent variables (factors). Experimental subjects saw all factor levels, but not all combinations, since one exposure to a sentence would certainly influence another.

Five usability sessions preceded the productivity experiment reported in this case study. Two subjects from the eventual pool of five professional translators participated in each usability session. The format of each usability session was the same as the productivity experiment, although new linguistic data was introduced in each session. The first author observed subjects while they worked and occasionally intervened to ask and answer questions. These sessions served as training for a productivity experiment and represented subjects' only exposure to Lilt. No subject accumulated more than six hours in Lilt during the usability sessions.

The assignment of subjects to conditions and data sets was randomized according to Table 1. Subjects completed the one-hour morning and afternoon sessions under time pressure, and were allowed an untimed break between sessions. For each subject, the timer was not started until he or she was satisfied with the CAT tool configuration. Subjects were permitted to configure Trados according to their own preferences, and were also allowed to use any external resources such as online dictionaries and concordances. After completing both sessions, subjects completed an anonymous exit questionnaire.

Prior to the experiment, subjects were instructed to translate as quickly and as accurately as they could. A

	UI	H
segments	80	125
words	980	1,132
repeat rate	15.3	19.5
TM size	9,612	12,824
exact %	6.3	15.3

Table 2: Gross statistics for the corpora in the productivity experiment. *TM size* is the number of translation units. *repeat rate* is the repetition rate of the source text. *exact %* is the exact match rate relative to the TM.

$100 award was offered to the translator who achieved the highest throughput in any condition. Subjects completed the experiment under time pressure. Time pressure isolates translation performance from reading comprehension [3] while eliciting a physiological reaction that may increase cognitive function [2].

Selection of Translators

We recruited five translators (three female) from e2f translations,[1] a translation agency with offices in France, San Jose, and Madagascar. The five subjects represented the entire in-house translation staff in the San Jose office. Subjects had between two and 15 years of professional translation experience.

We chose in-house translators for three reasons. First, we could control for many confounding factors (e.g., internet connection speed, time of day, computing hardware, etc.) while retaining a known and natural setting for the subjects. Second, we could observe the subjects and ask questions during the usability sessions. Finally, SDL Trados does not record timing information, so we needed to proctor the experiment. In the first few usability sessions, we found that self-reporting was unreliable, so we chose not to include remote translators.

Subjects reported at least two years of experience with Trados, the most common tool used at the agency.[2]

Linguistic Data

The early usability sessions showed that translators were routinely slower in the morning session than in the afternoon irrespective of translation condition. Interviews revealed that translators would perform the bulk of their terminology research and solve the hardest translation problems in the morning. In the afternoon, less time was typically spent on research. To mitigate this effect in the final productivity experiment, we chose different genres for the morning and afternoon sessions. Actual client data was selected.[3]

The first corpus contained UI strings from a software product. The second was a hotel chain loyalty brochure. Each dataset had an associated TM. Table 2 shows gross statistics. We attempted to match the datasets in terms of TM coverage, TM size, exact-match rate, source repetition rate [4], and perceived difficulty.

Exit interviews revealed that translators found the UI data to be easier than the hotel data. Timing results confirmed that all subjects achieved higher raw throughput on the UI data.

Translation Conditions (CAT Tools)

SDL Trados Studio 2011 was the baseline condition. The agency owns licenses for more recent versions, but 2011 is the principal version used in production. Translators were permitted to configure Trados according to their preferences. However, we observed that subjects did not enable AutoSuggest (predictive typing) and MT, relying almost entirely on the core TM functionality. When asked about this preference, all subjects said these features rarely benefited their work.

The current version of Lilt does not have any configuration settings.

[1] http://e2f.com

[2] Like many translation agencies, e2f uses a variety of CAT tools—including but not limited to memoQ, Wordfast, Memsource, and Transifex—in production to satisfy client requirements.

[3] The client data was no longer covered by non-disclosure agreements.

	UI			Hotel		
	Trados	Lilt		Trados	Lilt	
Average	813.4	991.3	+21.9%	723.1	821.1	+13.6%
Maximum	862.4	1,148.7	+33.2%	762.6	874.2	+14.6%

Table 3: Main results. Quality-adjusted throughput for the productivity experiment. Tables 4 and 5 show the per-subject raw scores.

		Tput	Quality
1	T	980.0	3.9
3	L	1069.1	3.9
4	L	1367.4	4.2
5	T	980.0	4.4

Table 4: UI data set per-subject results. Subject number, translation condition, raw throughput (*Tput*) in source words per hour, and raw quality scores. Subject 2 was excluded from the final analysis due to prior exposure to the client data.

		Tput	Quality
1	L	974.0	4.2
2	L	771.0	5.0
3	T	930.0	4.1
4	T	712.0	4.8
5	L	930.0	4.7

Table 5: Hotel data set per-subject results.

Evaluation Metrics

We recorded time for each session in order to compute throughput, which is source words per hour. We report average and maximum throughputs.

The optimal strategy for maximizing throughput in Lilt would be to accept the raw MT for every segment. To force subjects to trade-off between speed and accuracy, we submitted the final translations to e2f's production quality control process. The review was blind to subject and translation condition. One native French reviewer in Madagascar rated each document on a five-point scale for the following five attributes:

1. Style
2. Grammar
3. Source adequacy
4. Spelling and punctuation
5. Terminology

We averaged the five components to create a final quality score, also on a five-point scale. Multiplying this score by the raw throughput gives our main evaluation metric: **quality-adjusted throughput**, the unit of which is quality-adjusted words per hour.

Results and Analysis

Table 3 shows the main results for quality-adjusted throughput. For the UI data, subjects were 21.9% faster on average with Lilt. When comparing the fastest subject in each condition (Maximum), the difference is even more significant: a 33.2% margin. That fastest subject in Lilt also achieved the highest raw throughput in any condition at 1367 words per hour. Average quality was comparable in both conditions: 4.1 in Lilt and 4.2 in Trados. The agency typically releases translations to clients if the quality score is greater than 4.0.

As for raw throughput, subjects were 24.3% faster in Lilt on average. The fastest translator in Lilt was 39.5% faster than the fastest translator in Trados.

For the hotel data, subjects were still faster in the Lilt condition, but the differences were less significant: 13.6% on average, and 14.6% higher when comparing the fastest subjects. The average quality scores were once again comparable: 4.6 in Lilt and 4.5 in Trados.

The exit questionnaire affirmed the quantitative results. When asked to identify the translation condition that maximized throughput, three of the five subjects (60%) chose Lilt. When asked to what degree they agreed with the statement "In Lilt, the predictive translation suggestions increased my translation throughput," four

subjects selected “strongly agree,” and one selected “agree.” When asked to what degree they agreed with the statement “In Lilt, the predictive suggestions improved while I worked,” four of the subjects agreed.

Our previous work on translation productivity [5] used linear mixed-effects models to isolate the effect of translation condition from random variation due to subjects, genres, and source sentences. However, this analysis requires per-segment timings, which Trados does not record. Consequently, in this work we report averages and maximum scores in each condition. To compute statistical significance, the sample size should be greatly increased, or Trados should be instrumented to collect per-segment timings.

Conclusion and Future Work

Five in-house English–French translators translated UI and hotel data in Lilt and Trados under significant time pressure. The main findings are:

1. UI data: Lilt was 21.9% faster on average
2. Max raw throughput: 1,367 words / hour (Lilt)
3. Hotel data: Lilt was 13.6% faster on average
4. Final quality is comparable in the two tools

Recall that translators had less than six hours of experience in Lilt, and that baseline thoughput exceeded 5,000 words per day, a high rate relative to reported industry averages [7].

When asked to identify the single feature that could further improve productivity in Lilt, all subjects identified partial TM matching, a core feature of Trados.

The generality of these results may be limited by the subject sample size. If we need 16 subjects with per-segment timings—the setup in our previous work [5]—then we require either a very large in-house translation staff or remote translators. We are unaware of the former, and the latter would introduce additional confounding factors. Nevertheless, we are investigating the feasibility of large-scale experiments.

REFERENCES

1. P. J. Arthern. 1979. Machine Translation and Computer Terminology Systems: A Translator’s viewpoint. In *Translating and the Computer*, B.M. Snell (Ed.). North-Holland Publishing.
2. G. Bayer-Hohenwarter. 2009. Methodological reflections on the experimental design of time-pressure studies. *Across Languages and Cultures* 10, 2 (2009), 193–206.
3. S. Campbell. 1999. A cognitive approach to source text difficulty in translation. *Target* 11, 1 (1999), 33–63.
4. M. Cettolo, N. Bertoldi, and M. Federico. 2014. The Repetition Rate of Text as a Predictor of the Effectiveness of Machine Translation Adaptation. In *AMTA*.
5. S. Green, J. Chuang, J. Heer, and C. D. Manning. 2014. Predictive Translation Memory: A Mixed-Initiative System for Human Language Translation. In *UIST*.
6. E. Horvitz. 1999. Principles of Mixed-initiative User Interfaces. In *CHI*.
7. Rebecca Ray. 2013. Ten essential research findings for 2013. In *2013 Resource Directory & Index*. Multilingual.