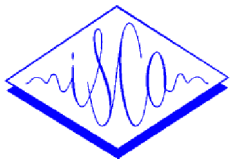
Proceedings of IWSLT 2008, Hawaii - U.S.A.



Evaluating Productivity Gains of Hybrid ASR-MT Systems for Translation Dictation

Alain Désilets [1], Marta Stojanovic [1], Jean-François Lapointe [1], Rick Rose [2], Aarthi Reddy [2]

[1] Institute for Information Technology, National Research Council of Canada
{alain.desilets, marta.stojanovic, jean-francois.lapointe}@nrc-cnrc.gc.ca

[2] Department of Electrical and Computer Engineering, McGill University
{rose, aarthi.reddy}@ece.mcgill.ca

Abstract

This paper is about *Translation Dictation with ASR*, that is, the use of Automatic Speech Recognition (ASR) by human translators, in order to dictate translations. We are particularly interested in the productivity gains that this could provide over conventional keyboard input, and ways in which such gains might be increased through a combination of ASR and Statistical Machine Translation (SMT). In this hybrid technology, the source language text is presented to both the human translator and a SMT system. The latter produces N-best translations hypotheses, which are then used to fine tune the ASR language model and vocabulary towards utterances which are probable translations of source text sentences. We conducted an ergonomic experiment with eight professional translators dictating into French, using a top of the line off-the-shelf ASR system (Dragon NatuallySpeaking 8). We found that the ASR system had an average Word Error Rate (WER) of 11.7%, and that translation using this system did not provide statistically significant productivity increases over keyboard input, when following the manufacturer recommended procedure for error correction. However, we found *indications* that, even in its current imperfect state, French ASR *might* be beneficial to translators who are already used to dictation (either with ASR or a dictaphone), but more focused experiments are needed to confirm this. We also found that dictation using an ASR with WER of 4% or less would have resulted in statistically significant ($p < 0.6$) productivity gains in the order of 25.1% to 44.9% Translated Words Per Minute. We also evaluated the extent to which the limited manufacturer provided Domain Adaptation features could be used to positively bias the ASR using SMT hypotheses. We found that the relative gains in WER were much lower than has been reported in the literature for tighter integration of SMT with ASR, pointing the advantages of tight integration approaches and the need for more research in that area.

1. Introduction

In the days before desktop computers and word processors, most professional translators used dictaphones. While this practice allowed them to very rapidly compose translations, it had one severe drawback, namely, clerical personnel had to later transcribe the audio to text. This turned out to be awkward and costly, and it is the main reason why the practice was dropped when desktop computers appeared in the eighties. With those new tools, translators could now directly type their own translations without having to go through an intermediary for transcription to text.

While this was hailed as an innovation by some, many translators felt a drop in their "personal" productivity (notwithstanding the transcription costs and delays) on account of their own slow typing speed, and the increased cognitive load of having to focus on both translating and typing. Consequently, even today, use of dictaphone and transcription staff is still practiced by some agencies, and anecdotal reports abound to the effect that this makes better use of the translator's time. Indeed, studies conducted 40 years ago claimed that the productivity of human translators *might be as much as four times higher* when dictating [1].

A priori, it may seem that modern Automatic Speech Recognition (ASR) provides translators with the best of both worlds. In principle, it could allow them to dictate translations directly into a word processor, without incurring the costs and delays of transcription by clerical staff. Indeed, the following quotes from two participants on the *SR_for_translators* mailing list are representative of anecdotal reports made by translators who are “enthusiastic” users of ASR:

> *“I'd guess that with the above setup [two monitors, Trados with capable Translation Memory, and a multi-button mouse], I'm three times more productive than I would be with just Trados and a single monitor. I'd guess that SR is 50% of that boost”*

> *“Speech recognition allows me to input two or three times as many words per hour”*

Increasing the productivity of translators by such large twofold factors or more would represent a significant economical impact, considering that the translation industry in Canada alone may approach 500 million dollars per year and may employ well over 13,500 translators [2]. In spite of those large potential benefits, ASR systems are rarely used in the translation industry today, even though modern off-the-shelf ASR systems routinely achieve recognition rates in the order of 95% or more for English. While this may seem like it would be more than adequate, in practice people who try ASR systems (whether for a translation task or not) often report that an error rate of 5% is still unacceptable.

This may come as a surprise, but one must realize that in the context of dictation with an ASR system, correcting a single transcription error can easily require 15 to 30 seconds.

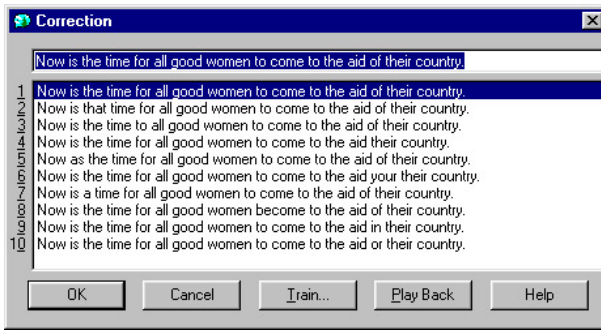


Figure 1: The Dragon Naturally Speaking error correction dialog

Typically, a user will notice a text segment that makes no sense, and conclude that it was partly mistranscribed. In the specific case of the Dragon Naturally Speaking ASR system, the user would then have to select that segment with the mouse, and utter a *"correct that"* command, in order to start the correction dialog (see Figure 1). Often, the user cannot remember exactly what it was that he said, and he must utter a *"play that back"* command, then listen to a play back of his own speech. Once the user knows exactly what he said, he is ready to do the correction. Typically, he will start by scanning the list of suggested corrections to see if it includes the correct transcription. In our experience, more often than not the correct transcription is not in the list of suggestions, and the user must click on the best suggestion, then fix it with mouse and keyboard before hitting the *OK* button.

Note that while the above discussion is based on the Dragon Naturally Speaking error correction procedure and user interface, it is very similar to that of other commercial ASR systems like the one found in Windows Vista. Also, ASR manufacturers recommend that users diligently correct each and every recognition error in that way, so that the ASR can learn from its mistakes and improve its accuracy. With some ASR systems, failing to correct recognition errors may even result in a progressive degradation of accuracy, because the system performs continuous adaptation based on the user's speech, under the assumption that the transcription is accurate unless the user says otherwise.

In short, the cost of individual transcription errors is clearly high, and any improvement in ASR accuracy, even a few percentage points, may thus greatly enhance usefulness of this technology and lead to the realization of the productivity gains and economic impacts described above for the translation industry.

In this paper, we specifically investigate the following two research questions:

- **Question 1:** Are current off-the-shelf commercial ASR systems sufficiently accurate to provide a productivity gain for professional translators? And if so, what is the order of magnitude of that gain?

- **Question 2:** Can the productivity gains be increased by combining ASR with Statistical Machine Translation (SMT), in such a way that the SMT system provides hints to the ASR system as to what the translator is likely to utter when translating a particular source text?

The SMT and ASR combination mentioned in Question 2 requires a bit of explanation. In this hybrid technology, the source language text is presented to both the human translator, and a SMT system (Figure 2b). The later then produces N-best translations hypotheses which are used to fine tune the ASR language model and vocabulary, towards utterances which are probable translations of source text sentences.

Note that *Spoken Language Translation* (SLT) [3] is a similar research field that also involves hybrid SMT and ASR technologies, however its paradigm for combining them is quite different. In Spoken Language Translation, the aim is to take spoken audio in a source language and transform it to written text or spoken audio in a *different target language* (Figure 2a). It is worth noting that Translation Dictation with ASR is a much easier task than Spoken Language Translation. Indeed, in Translation Dictation with ASR, SMT and ASR systems are combined in such a way that their respective errors hopefully cancel each other. In contrast, in a typical Spoken Language Translation situation, the errors of both systems tend to multiply each other. One exception is the SLT scenario depicted in Figure 2c, where spoken audio is uttered by a speaker in source language, and its simultaneous translation is uttered in parallel by an interpretor. In such a scenario, accuracy of SLT for the original source language track can be improved by leveraging clues obtained by carrying out SLT on the simultaneous translation track [4].

The remainder of the paper is organized as follows. Section 2 describes related work. Section 3 describes an experiment to collect audio and productivity data from professional translators using keyboard and ASR. Section 4

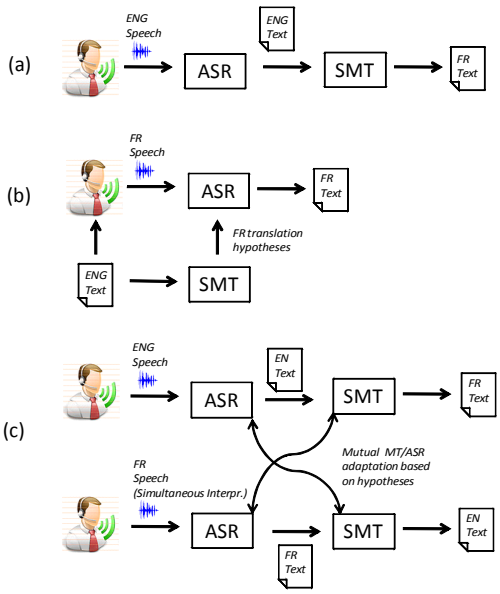


Figure 2: Data flow for an English to French translation scenario, with: (a) Spoken Language Translation, (b) Translation Dictation with ASR, and (c) Spoken Language Translation with simultaneous interpretation audio track.

uses this data to evaluate the potential productivity increase of ASR over keyboard. Section 5 uses the same data to evaluate different strategies for combining SMT and ASR. Finally, Section 6 summarizes the conclusions.

2. Related work

To our knowledge, the only formal evaluation of the productivity of translation dictation is a 1963 study by Sinaiko [5]. The famous ALPAC report [1] cites it as follows:

> *"One experiment that has come to the attention of the Committee indicates that a rapidly dictated translation is almost as good as a "full translation", and* ***takes only about one fourth the time****"*

However, looking closer at the actual Sinaiko study, it is clear that this result is not directly applicable when trying to evaluate the productivity of translating with ASR versus keyboard. Firstly, the two conditions investigated, called *sight translation* and *full translation,* differed in more ways that the input modality. Indeed, the full translation condition involved an additional review step, which the sight translation condition did not incur. Given that revising a translation tends to be much faster than initial translation, one might still surmise that most of the productivity gain is indeed attributable to the difference between the dictation versus writing modalities. However, a second issue is raised by the fact that the full translation teams are described as having had a typist at their disposal, which seems to indicate that the translators might not have typed their translation themselves. It could be that they were writing their translations with pen and paper (which is bound to be slower than typing with modern word processing technology), or that they were dictating them to a tape recorder or to the typist (in which case, the productivity gain of sight translation cannot be attributed to dictation). A third issue is the fact that the two conditions involved very different kinds of subjects. In the sight translation condition, subjects were interpreters, who are used to translate verbally in real time, and it is not clear that "normal" translators would be able to dictate translations as fast without extensive interpreter training. Even notwithstanding those three issues, results of this study do not apply directly to an ASR scenario, because the sight dictation condition involved dictation to a human typist, as opposed to an ASR system. Therefore, the study does not account for the cost of correcting ASR recognition errors. In contrast, the study described in our paper does not present any of the above limitations. To our knowledge, it is the first controlled investigation of the possible productivity gains of Translation Dictation with ASR.

Another goal of the present paper is to evaluate how productivity gains might be increased using various approaches for combining SMT with ASR. One of the earliest efforts made in that direction was performed by Brown et al. [6]. In that work, the optimum target language string in a stack decoder based ASR system was obtained from the joint probability of the source language and target language strings computed from both Language Model (LM) and translation model parameters. While they did not report ASR results, they demonstrated that the perplexity of the combined translation/language model was significantly less than the original trigram LM for utterances taken from the Canadian Hansard corpus.

At about the same time, Brousseau et al. [7] presented two methods for combining ASR and SMT models as part of the TransTalk project which involved English to French translation in the Canadian Hansard domain. Of particular interest was a method for re-scoring N -best lists of French word hypotheses generated by a large vocabulary continuous speech recognizer (LVCSR) with a language translation model.

More recently, Paulik et al. [8] applied a number of techniques to achieve a closer integration between text based SMT and acoustic ASR on an English to Spanish travel-phrase language translation task. Integration was accomplished both through rescoring of N -best word hypotheses and by incorporating candidates from SMT into cache and interpolated LMs for ASR. Khadivi et al. [9], demonstrated a decrease in WER on an English-German technical document translation task when using different translation models to re-score the N -best lists obtained from the recognizer. An interesting result of both of these recent papers was that the largest increase in ASR performance was obtained not from the very best performing translation models, but instead from translation information that incorporated limited word context information. Khadivi et al [10] also investigated several tighter integration techniques for combining the word graphs of the ASR and SMT systems in an English-German translation task. They found the results of these approaches to be similar to that of N-best rescoring, especially for larger values of N.

Reddy et al. [11] described two methods of integrating ASR and SMT systems. In one method, called *loose integration* the target language N-gram LMs generated by the SMT are combined with the LM used in the ASR system by either interpolating the ASR LM or by rescoring with the SMT-LM after ASR decoding. In the other method called *tight integration*, the SMT translation models combined with ASR LM and acoustic scores are used to improve ASR lattice decoding. The authors found that WER improvements were substantially larger with tight integration than with loose integration. The hybrid ASR-SMT strategies described in that work were based on PORTAGE, a phrase based SMT system developed at the NRC Institute for Information Technology [12]. The same PORTAGE SMT system is used in the present paper.

Although technically focused on an SLT task, the work of Paulik and Waibel [4] is relevant for Translation Dictation with ASR. Here, accuracy of SLT is carried out simultaneously on an audio track uttered in source language, and its simultaneous translation uttered in parallel by an interpreter in the target language. N-grams and full translation hypotheses are produced for each of the two SLT tasks, and are used to bias both the ASR and MT steps of the other SLT task. Using those strategies, the authors find small but consistent BLEU score improvements.

It should be noted that none of the previous work on hybrid ASR-SMT systems evaluated the impact on actual *translator productivity*. In particular, none of these studies compared productivity of the hybrid system to a traditional desktop word processing situation. In contrast, the present work evaluates the impact in terms of a productivity measure called *Translated Words Per Minute* (*TWPM*), and compares

	ASR first	**ASR second**
ASR used for ST1	2 subjects	2 subjects
ASR used for ST2	2 subjects	2 subjects

Table 2.1: Experimental design for data collection.

productivity with a traditional desktop word processing situation.

3. Experimental design for data collection

In order to investigate our two research questions, we collected audio and productivity data from eight professional translators. In this section, we describe the experimental design for this data collection experiment.

All subjects were professional translators working in the English to French direction. Translation into French (as opposed to English) was chosen because it corresponds to the situation that strongly predominates in the Canadian translation industry. It also reflects the reality of other contexts (ex: the European Union) where English is seldom the target language for translation. This turns out to be important for our evaluation because current ASR technology is much more advanced for English than for other languages. Six of the subjects were senior translators with at least 15 years of work experience, and two were junior translators with less than five years. Only one subject was already using ASR to dictate translations, on account of a Repetitive Strain Injury that limited the amount of daily typing she could do, but did not affect her typing speed. Another subject used a dictaphone as her regular mode of translation, while the remaining six subjects used keyboard. Two of those six had previously used dictaphones for at least 10 years, but had not used this mode in recent years. Another two of those six had tried commercial ASR, but did not end up adopting it in their work practice. The last two subjects had never tried either ASR or dictaphone.

In order to factor out a possible effect of subject variability, we used a within group experimental design, where each subject translated one text using ASR and one text using mouse and keyboard. Both source texts (respectively referred to as *ST1* and *ST2*) were taken from the Canadian Hansard (i.e., transcriptions of debates at the Canadian House of Commons), and both were on the same topic, namely, involvement of Canadian troops in the Irak war. This topic was chosen because we could assume that all subjects were familiar with it. In order to factor out possible effects due to the difference of source texts, half of the subjects translated ST1 with ASR, and ST2 with keyboard, while the other half did the opposite (i.e. ST1 with keyboard and ST2 with ASR). Similarly, in order to avoid possible effects due to increased familiarity with the source text topics, half of the subjects first did the ASR task, followed by the keyboard task, while the other half started with the keyboard task, followed by ASR. In summary, subjects fell in one of 4 cells shown in Table 2.1. Although typing speed and a-pirori familiarity with ASR and dictation technologies were measured, we did not control for those variables in our data collection protocol.

The ASR system used was the French version of Dragon Naturally Speaking 8. Before carrying out the various evaluation tasks, subjects were asked to do a number of preparatory tasks to train this ASR system and familiarize themselves with its use. First, the subject carried out the standard enrollment procedure required by the system, under the supervision of a researcher. This consisted of reading certain documents out loud to the ASR, so that it could adapt to the subject's voice. The researcher then gave a 15 minutes demo of the ASR system, and how to use it to dictate translations. In particular, subjects were advised to compose full sentences or paragraphs before proceeding with correction, and also, to look away from the screen while dictating. The latter is a common best practice for dictation with ASR, because transcription errors appearing in real time on the screen can seriously distract the user and disrupt his flow of thought and speech.

The demo was followed by a 15 minutes practice session where the subject used the ASR system to dictate the translation of a text similar to ST1 and ST2 (namely, a text from the previous Hansard day, also on the topic of Canadian involvement in the war in Iraq).

After this training period, the subject was asked to carry out terminology and phraseology searches for both ST1 and ST2 texts, before starting translation. This is a best practice for translation in general, but one that is particularly important in a dictation context. Indeed, translators who dictate (either with ASR or a dictaphone) often report that interruptions in flow caused by terminology or phraseology difficulties tend to be more disruptive when dictating than typing.

While the subject was doing terminology searches, the experimenter carried out Domain Adaptation (DA) using the facilities provided by the manufacturer of Dragon Naturally Speaking. In total, we poured 3.9 million words of text, into the domain adaptation module. These consisted of Hansard transcriptions for the 6 months that immediately preceded (but did not include) the day from which the test texts ST1 and ST2 were taken. This adaptation required approximately 30 minutes of processing time on the computer.

After these preliminary searches, the subject carried out the two translation tasks using the order and source texts prescribed by their assigned cell in Table 2.1. In both cases, the subject translated until task completion, or until 30 minutes had elapsed. In both cases, the subject was asked to aim for a first draft version of the translation. This was done in order to take revision and reformulation time out of the equation. At the end, a short debriefing interview was carried out to discuss what the subject liked and disliked about the ASR system.

The following data was collected:

- Enrollment audio
- Audio from ASR task
- Screen capture of the ASR and the Keyboard tasks
- Text from the translations as they stood at the end of each task.
- Audio from the debriefing interview

The enrollment audio data consisted of approximately six minutes of audio spoken in French by each of the speakers, amounting to a total of 49 minutes of speech. The audio from the ASR task consisted of 5,748 words spoken in French, amounting to a total of 81 minutes of speech (after removal of long silent pauses during which the translator was

reflecting). All audio was recorded using a Shure headset microphone with external soundcard pod.

4. Productivity analysis

4.1. Performance measures

In this study, we used two measures of performance: Word Error Rate (WER) and Translated Words Per Minute (TWPM). WER is the standard measure used for Speech Recognition, while we define TWPM as follows:

$$TWPM = W / T \quad (1)$$

where *W* is the number of source text words that where actually translated, and *T* is the total task time (in minutes). In the case of an ASR task, *T* can further be decomposed as follows:

$$T = D + C \quad (2)$$

where *D* is the time spent actually dictating translations (including any time spent reflecting), and *C* is time spent correcting speech recognition errors. Using the screen capture from the ASR task, we were able to count *D* and *C* time separately.

Table 4.1 summarizes the values of these two measures for a series of conditions which are described and discussed in the remainder of section 4 and in section 5.

4.2. Productivity of ASR versus keyboard

The bar graph in Figure 3 graphically shows the average TWPM for various conditions (taken from Table 4.1).

The *Keyboard* and *ASRBaseline* bars display the average TWPM achieved by our subjects during the data collection experiment, when translating with keyboard and ASR respectively. Thus, the *ASRBaseline* includes Domain Adaptation (DA) based on the 3.9 million words of Hansard text. In the *ASRAdvUser* condition, we simulated what would have happened if all subjects had been experienced ASR users with proper mastery of the error correction procedure. We calculated the average *C* time per word corrected, for the one subject who had been using Dragon NaturallySpeaking for years in her daily translation work. We then applied this correction speed to adjust the *C* time for all other subjects accordingly.

The *ASRPerfect* bar corresponds to the average TWPM that our subjects would have achieved if they had used a "perfect" ASR system with WER = 0. This TWPM was calculated by removing the *C* time from the task time calculated for the *ASRBaseline* condition. The remaining bars display TWPMs for simulated ASRs with increasing accuracies. These were obtained by downscaling the *ASRBaseline* WER of each subject by a constant factor $0 < r < 1$. We then computed the correction time that a subject would have experienced, by assuming that it was directly proportional to the WER. In other words, for a given *WER'*, we assumed that the corresponding correction time *C'* was:

$$C' = C_B [1 - (WER_B - WER') / WER_B] \quad (3)$$

	TWPM	WER
Keyboard	20.7 (1.6)	N/A
ASRBaseline	20.5 (2.7)	11.7 (1.6)
ASRAdvUsers	23.0 (2.3)	11.7 (1.6)
ASR 5%	24.9 (3.0)	5.0 (0.7)
ASR 4%	25.9 (3.0)	4.0 (0.6)
ASR 3%	26.7% (3.0)	3.0 (0.4)
ASR 2%	27.7% (3.0)	2.0 (0.3)
ASR 1%	28.8 (3.1)	1.0 (0.1)
ASRPerfect	30.0 (3.1)	0.0 (0.0)
NoDA	20.0 (2.8)	12.9 (2.0)
100BestSMTx1	20.5 (2.8)	12.0 (1.9)
100BestSMTx50	20.4 (2.8)	12.0 (1.8)
100Bestx50+Hans	20.8 (2.8)	11.5 (1.8)

Table 4.1: Average TWPM and WER for different conditions (standard deviation in parentheses).

where WER_B and C_B correspond to values for the *ASRBaseline* condition. Note that this interpolation scheme assumes that *relative* improvements are the same for all subjects. This would be obviously wrong for *absolute* improvements, because those tend to be larger for subjects whose WER is large to start with. But relative improvements are normalized for scale and may tend to be more consistent across subjects. Also, the interpolation scheme does not take into account the fact that certain types of errors (ex: plural forms) require less time than others for the human translator to fix. Because of those limitations, our interpolated TWPM should only be taken as indicative of the productivity that translators would have experienced at given WER levels. Note however that in the case of *ASRPerfect*, no interpolation was used, and that it can be therefore interepreted as an accurate reflection of the productivity which our subjects would have *actually* experienced with a perfect ASR system.

Each of the *ASR X% WER* (X = 5, 4, 3, 2, 1) thus corresponds to a simulated ASR whose WER would be *X%* . The X=5 scenario corresponds to a conservative assessment of the average WER for recent off-the-shelf English ASRs (which are generally much better than French ASRs), and X=1,2 corresponds to manufacturer-claimed accuracy for the latest English version of Dragon Naturally Speaking (version 9).

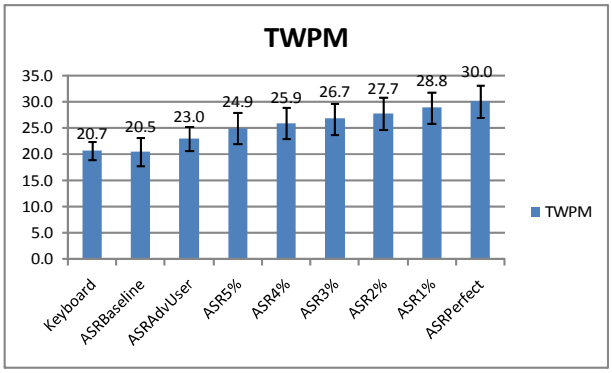


Figure 3: Translated Words Per Minute (TWPM) for various scenarios.

ANOVA (analysis of variance) reveals that the differences between these various conditions are statistically significant ($p < 0.01$). However, the *Keyboard*, *ASRBaseline* and *ASRAdvUser* conditions did not significantly differ. In other words, even correcting for our subjects' lack of experience with the ASR error correction procedure, we did not find dictation with ASR to be better than with keyboard. However, we found that the simulated ASRs with WER $\leq$ 4% were significantly different from the *Keyboard* condition ($p \leq 0.06$).

Although these results cannot be used to conclude that translating with current commercial French ASR is no better than with keyboard, it does shed serious doubt on the very large productivity gains that are being reported anecdotally. If, as is often claimed, translating with ASR was truly twice as fast as with the keyboard, a within-group experiment should have been able to show a statistical difference even with only eight subjects. On the other hand, these results clearly indicate that dictating translations using a higher accuracy ASR with WER $\leq$ 4% would result in substantial productivity gains in the order of 25.1% to 44.9% TWPM. But again, even this falls quite short of a twofold productivity increase.

It is worth noting that our data is mostly representative of the performance experience initially, by translators who are not used to dictating. Indeed, only two of our subjects used dictation in their regular translation practice (one with ASR, the other with dictaphone). This does not diminish the importance of our findings, because a positive initial user experience is vital for ASR adoption. Many first time users end up abandoning that technology if they do not find it useful after a few days of use. However, this does beg the question of whether our subjects might have experienced productivity gains, once they got used to dictating. Looking more closely at the two subjects who used dictation in their regular practice, we find that they respectively experienced relative gains of 34.8% and 37.8% in TWPM. In contrast, only one of the remaining six subjects experienced a productivity gain, and it was much smaller, namely, 17.6%. This seems to indicate that even in its current imperfect state, French ASR might indeed benefit translators who are already used to dictating. However, experiments with more subjects of that type would be needed to confirm this. In particular, it could be that there is an implicit selection bias, in that translators who have already adopted dictation in their regular practice, did so because of an innate ability for dictating. It is not clear that other translators would start out with that particular skill, nor that they could acquire it rapidly enough to make adoption of ASR palatable. It is also worth pointing out that even the gains for those two dictation-experienced subjects still fall short of the large twofold or more increases being reported anecdotally.

Note also that productivity statistics may not tell the whole story. Indeed, in the debriefing interviews, more than half of the subjects said they enjoyed their ASR experience, and that they would seriously consider using it for their actual work. This, in spite of the fact that they had a fairly realistic assessment of their productivity gain (or lack thereof) with ASR. Subjects who said they would consider using the ASR seemed to feel that dictation put them in a different mental mode, which they thought was more conducive to good translations. More research needs to be done to evaluate more precisely user's subjective perception of the experience, and whether that perception remains after the initial "honeymoon" period.

5. Productivity for various SMT and ASR combinations

Having established that current commercial French ASR technology falls short of achieving productivity gains for translators, we now investigate the extent to which combinations of SMT and ASR might bridge that gap.

We evaluated several strategies for integrating SMT hypotheses, using the limited manufacturer provided features for Domain Adaptation (DA). All variants included acoustic adaptation based on the subjects' audio enrollment data.

We used two ASR variants as baselines. The first one is the *ASRBaseline*, described earlier, which incorporates DA based on 3.9 million words of Hansard text. The second baseline, *NoDA,* corresponds to a strategy where no DA is carried out at all. The reason we investigated this approach is that the *ASRBaseline* variant assumes the translator has access to 3.9 million words of text in the target language, similar to the source text she is about to dictate. But most translators may not have access to that much relevant training text.

In the *100BestSMTx1* variant, we carried out DA based on the 100 best translations proposed by the SMT system for

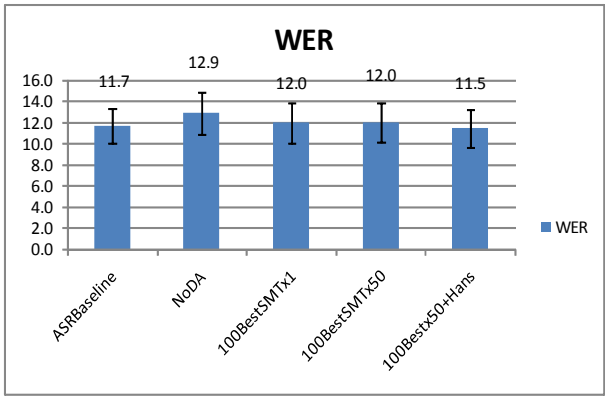


Figure 4: Word Error Rate (WER) for different ASR variants.

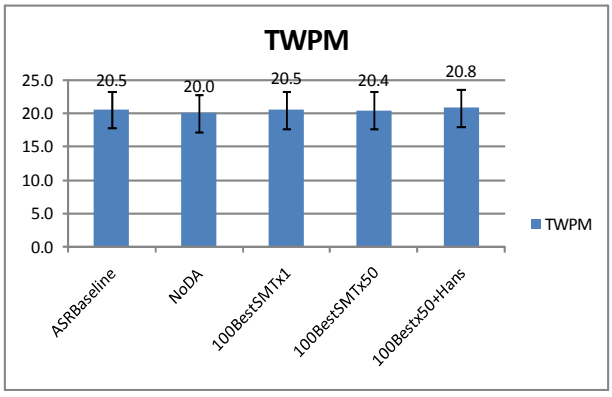


Figure 5: Translated Words Per Minute (TWPM) for different ASR variants.

each of the source text sentences being translated by the subject. This amounted to an average of 76,048 words for ST1 and ST2. DA based on Hansard text was not used. The purpose of this variant is to evaluate the degree to which DA based on a relatively small but highly targeted corpus generated by the SMT, can achieve the same kinds of improvements as DA based on a much larger but more generic corpus extracted from translation archives. This has practical application since DA based on the 76,048 words from the SMT only took 17 seconds, which is fast enough to envisage doing this on-the-fly to prepare the ASR for the translation of a specific source text.

The *100BestSMTx50* variant is similar to *100BestSMTx1*, except that DA was done using 50 copies of the 100 best translations. Multiple copies were used to simulate weighted LM interpolation, as in Reddy et al. [11]. This was necessary since Dragon NaturallySpeaking does not provide the user with a way to directly control these weights. The number of copies was calculated so that it would result in about the same amount of text as the 3.9 million words of Hansard used in *ASRBaseline*.

Finally, *100Bestx50+Hans* is similar to *100BestSMTx50*, except that we carried out DA based on the 3.9 million words of Hansard, in addtion to DA based on SMT outputs.

Note that all those variants only use what Reddy et al. called loose integration. While tight integration (ex: lattice re-scoring) might have provided higher accuracy gains, we were not able to experiment with it because the commercial off the shelf ASR system we employed (Dragon Naturally Speaking) did not provide necessary hooks to do so.

Figure 4 shows the WER for the ASR variants described above, on audio collected during the experiment with human subjects. Figure 5 shows the TWPMs, which were computed based on WER using formula (3). Both those figures are graphical representations of the figures in Table 4.1.

ANOVA analysis of the WER figures shows significant differences between the ASR variants ($p < 0.01$). However, while the two variants that included Hansard DA were statistically different from *NoDA* ($p < 0.01$), this was not the case for the two variants based on SMT DA only. Moreover, the two SMT only variants did not differ significantly from the two variants that included Hansard DA. In other words, SMT-based DA does not improve WER, whether done by itself, or in addition to Hansard DA. ANOVA analysis of TWPM reveals the exact same trends.

Note that these results should not be interpreted as meaning that a combination of ASR and SMT could not potentially improve WER and TWPM, because in our experiment, we were only able to use the very limited manufacturer provided features for DA. Given the small size of the texts used (7,028 words), it is not clear that these features allowed us to push hard enough on the system's internal LM (even when pouring 50 copies of them). Indeed, the best relative WER improvements we found (1.4% over *ASRBaseline*) are much smaller than those found by Reddy et al using tighter SMT-ASR integration (18.2% over a comparable baseline).

Although we were not able to evaluate productivity gains of a tight integration approach, we can hypothesize that it would result in the same relative improvements found by Reddy et al. This is reasonably sound, since the two studies are relatively comparable. Indeed, they both used the exact same experimental design and the exact same source texts (although in opposite translation direction), and the translation hypotheses were produced by the same SMT system (PORTAGE). The main differences between that study and the present one are:

- different translation directions (Fr->En versus En->Fr)
- different number of subjects (three versus eight).
- different ASR systems (research ASR system [13] versus Dragon Naturally Speaking)
- different amount of Hansard text used for standard adaptation (Reddy et al. used twice as much text, which included the text used in our study).
- different baseline WERs (20.8% versus 11.7%).

The Reddy et al. study found a 18.2% relative improvement in WER for tight integration using lattice rescoring, over a baseline with acoustic and standard LMA based on Hansard archives. Applying this improvement to our loosely equivalent *ASRBaseline* yields an average WER=9.6%, which we found to be statistically better than *ASRBenchmark*. However, using equation (3) to interpolate correction time, we find that this would result in an average TWPM=21.7, which amounts only to a 4.8% relative improvement over the *Keyboard* condition, and this was not found to be statistically significant. In other words, while tight coupling might have significantly improved WER, it would still have fallen short of improving productivity over keyboard, in the French dictation scenario we evaluated.

6. Conclusions and Future Work

Our study provides strong evidence that current commercial French ASR systems fall short of the very large twofold or more productivity increases being reported anecdotally by some translators. Indeed, at a baseline 11.7% WER, we found no statistically significant productivity improvement over keyboard, even when adjusting for our subject's lack of familiarity and skill with the manufacturer recommended error correction procedure. However, we did find *indications* that, even in its current imperfect state, French ASR *might be* beneficial for translators who are already used to dictation, but more experiments with users of that type needs to be carried out to confirm this. In any case, even for the two subjects who fell in that category, we found that the productivity gains still fell short of anecdotal reports (34.8% and 37.% Translated Words per Minute respectively).

On the other hand, we found that translators using better ASR systems with WER of 4% or less (which is well within the range of English commercial ASR systems) would experience statistically significant productivity gains in the order of 25.1% to 44.9% Translated Words per Minute. But again, this still falls short of a twofold increase, especially considering that time spent doing terminology searches and revising the first draft of the translation were excluded from the study. Since together these two activities often take as much time as the actual composition of the first draft, we might expect relative gains for a global translation task to be even smaller.

Our debriefing interviews also indicate that productivity statistics may not tell the whole story. Indeed, more than half

of our subjects had a positive impression of the system after using it, even though they had a realistic assessment of the productivity gain (or lack thereof).

Note that the usage scenario we evaluated assumes that the translator is responsible not only for dictating translations, but also for correcting transcription errors. Another model would be for the translator to focus only on dictation, and leave error correction to lesser paid clerical personnel. This is similar to what is being done in translation agencies that use dictaphone. In a scenario like this, we predict (based on *ASRPerfect*) that the translator himself would experience a 44.9% productivity increase for the creation of an initial draft. An interesting question in this scenario, is whether this would come at a price of decreased productivity for clerical staff, but unpublished work done at the Centre de Recherche Informatique de Montréal (CRIM) leads us to believe that this would not be the case [14]. The study compared the performance of humans transcribing audio from scratch, versus correcting errors in a draft transcription produced by ASR. It found the breakeven point to be around WER = 20%, which is well within the reach of any modern ASR systems.

Another usage scenario would be for the translator to overwrite errors directly with the keyboard, without going through the ASR's laborious error correction dialog. While this would make error correction faster, our data does not allow us to say by how much, and whether this would result in productivity increases over keyboard. Note however that in this scenario, the translator forgoes any possibility of the ASR improving its performance based on user correction, and that with some ASR systems, accuracy could even degrade as the ASR continuously adapts based on partially erroneous transcripts.

We also show that the limited Domain Adaptation features typically provided by commercial off-the-shelf ASR systems, are not sufficient to allow improvement of ASR accuracy based on SMT hypotheses. We found that the improvements that resulted from carrying DA with the top 100 best translation hypothesis were much smaller than what has been reported in the literature for tighter SMT-ASR integration on a comparable task, and that they turned out to not be statistically significant. This points out the advantages of tighter integration, and the need for more research in that vein.

7. Acknowledgements

The authors would like to thank the following people for their help. From NRC: George Foster, Roland Kuhn, Samuel Larkin, Pierre Isabelle, Julie Cliffe and Norm Vinson. From the Translation Bureau of Canada: Susanne Marceau and Susanne Garceau. Also, the eight anonymous subjects who participated in this study, as well as the two anonymous reviewers whose relevant comments greatly helped improve the paper.

8. References

[1] Pierce, J. R., Carroll, J. B., et al., “Language and Machines”, National Academy of Sciences, 1966.

[2] Duchaine, M., “Financing the language industry in Canada”, AILIA , 2006.

[3] Fordyce, C. S., “Overview of the IWSLT 2007 Evaluation Campaign”, 2007

[4] Paulik, M., Waibel, A., “Extracting Clues from Human Interpreter Speech for Spoken Language Translation”, 2008.

[5] Sinaiko, H. W., “Teleconferencing: Preliminary Experiments”, Institute for Defence Analyses, Alexandria, VA, USA, Report Number RP-P108 OR IDA-H-64-23 1963

[6] Brown, P. F., Chen, S. F., Pietra, S. A. S., Pietra, V. D., Kehler, A. S., and Mercer, R. L., “Automatic Speech Recognition in Machine Aided Translation”, Computer Speech and Language, 1994

[7] J. Brousseau, D. Drouin, G. Foster, P. Isabelle, R. Khn, Y. Normandin, P. Plamondon, “French Speech Recognition in an Automatic Dictation System for Translators: The TransTalk Project”, EuroSpeech'95, 1995

[8] Paulik, M. Fugen, C., Stuker, S., Schulz, T. Schaaf, T., Waibel, A., “Document Driven Machine Translation Enhanced ASR”, InterSpeech 2005.

[9] Khadivi, S., Zolnay, and Ney, H., “Automatic text dictation in Computer Assisted Translation”, InterSpeech 2005.

[10] Khadivi, S., Zens, R., Ney, H. “Integration of Speech to Computer-Assisted Translation using Finite-State Automata”. In Proc. COLING/ACL 2006, Sydney, Australia, July 2006.

[11] A. Reddy, R. Rose, A. Désilets, “Integration of ASR and Machine Translation Models in a Document Translation Task”, InterSpeech 2007, Antwerp, Belgium, Aug 27-31, 20007.

[12] Ueffing, N., Simard, M., Larkin, S., Johnson, J. H., “NRC's PORTAGE system for WMT 2007”,ACL-2007 Workshop on SMT, Prague, Czech Republic 2007

[13] Mohri, M., Pereira, F., Riley, M. .”Weighted Automata in Text and Speech Processing.”, 12th biennial European Conference on Articial Intelligence (ECAI-96), Workshop on Extended Finite State Models of Language, 1996.

[14] “Automatic Speech Recognition and Universal Accessibility of Parliamentary Debates and Committee Evidence (RAP)”, Canarie Workshop 2004, October 14, 2004, Vancouver, British-Columbia, Canada. http://www.canarie.ca/conferences/fall_series/vancouver/ppt/rap.ppt