



**NRC-CMRC**

*Institute for  
Information  
Technology*

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

**Alain Désilets**

**Marta Stojanovic**

**Jean-François Lapointe**

**National Research Council of  
Canada**

**Richard Rose**

**Aarthi Reddy**

**McGill University**



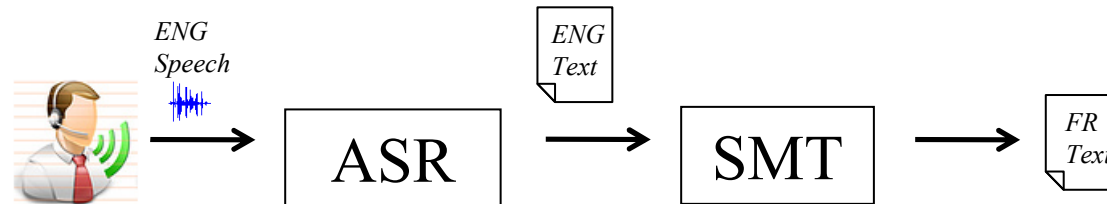
**National Research  
Council Canada**

**Conseil national  
de recherches Canada**

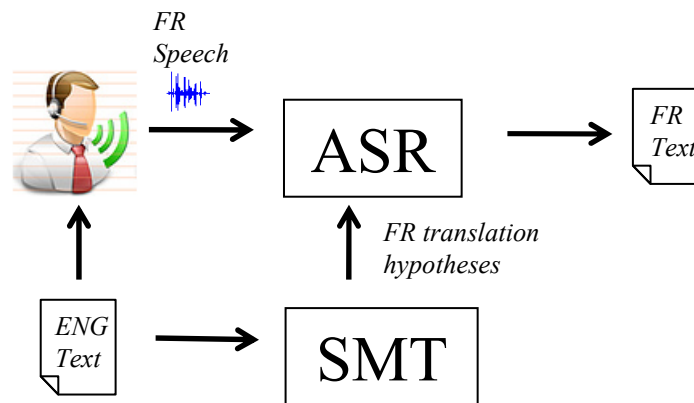
**Canada**

## Spoken Language Translation (SLT) vs Translation Dictation with ASR

SLT



Translation  
Dictation with  
ASR



## Spoken Language Translation (SLT) vs Translation Dictation with ASR (2)

For example, say a French source sentence is:

*“Je suis en avion en ce moment” (i.e., “I am in a plane right now”)*

Translator utters:

*“I’m in **a plane** right now”*

Which is one of the ASR’s hypothesis, but it thinks it might also have heard:

*“I am in **pain** right now”*

SMT gives high very high probability to word “*a plane*”, given that the source sentence contains French word “*avion*”. Can the ASR system exploit that information?

## Why is this important?

Translation industry is a \$ 500 million affair in Canada alone!

*Anecdotal reports* that translators dictating are MUCH MORE productive than those typing.

*“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”*

-- Posted on the SR for Translators mailing list

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

-- Posted on the SR for Translators mailing list

## Why is this important (2)?

*“My first boss translated using a dictaphone, and he **could keep three human transcribers busy full time with his recordings alone.**”*

*-- Anecdote collected from a translation professor*

*“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**“*

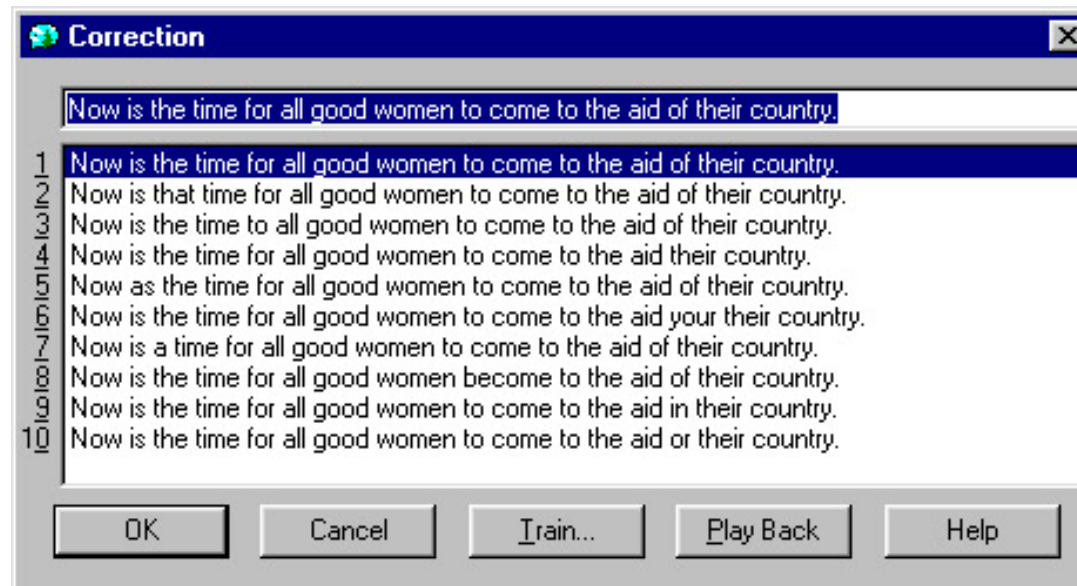
*-- ALPAC report, 1966*

Is this real, or just reports from over-enthusiastic users of SR, or confounded with other variables (ex: experience of the translator)?

## Is this real?

Users of desktop ASR often report that dictation is less productive than keyboard, even with accuracy as high as the lower to mid 90%.

Reason: Error correction is a very time consuming task!!!  
Fixing a single error can easily require 15 to 30 seconds.



# 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?

## Related work

### **Question 1: Productivity gains of off the shelf ASR systems for translation dictation**

Sinaiko et al., 1963 (cited in ALPAC), is only study, but:

- Description of experiment so vague to be meaningless
- Not done in an ASR context



## Related work (2)

### Question 2: Productivity gains of ASR+SMT combinations

Much work done since 1994:

- Brown et al., 1994, Brousseau et al., 1995, Paulik et al., 2005, Kadivi et. al, 2005, Reddy et al., 2007
- Different hybridisation strategies used:
  - Language Model Interpolation (LMI), N-best list re-scoring, word-lattice integration (tight integration)
- Clearly demonstrated improvements in WER
- **But is it enough to improve productivity of translators?**

# Data Collection Protocol

Conducted experiment with 8 professional translators to collect productivity and audio data.

## Subjects characteristics:

- Six with >15 yrs expérience, 2 with < 5 yrs.
- 1 using ASR in her work, due to Repetitive Strain Injury (Note: did not impede her typing speed).
- 1 using dictaphone in her work.
- 2 had used dictaphones during several years in the past.
- 2 had tried ASR before, but not adopted it.

## Data Collection Protocol (2)

Translation task in English->French direction

- Most representative of Canadian, and international situation (English rarely the target language for translation).

Intra-subject protocol

- each subject translated two documents (ST1 and ST2)
- one text translated with ASR, the other with mouse and keyboard.
- text of comparable difficulty and length.
- texts on same subject (Canadian Hansard debate on the involvement of Canadian troupes in Iraq).



## Data Collection Protocol (3)

Precautions taken to avoid systematic bias due to choice of text, and order of input modalities.

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

ASR was Dragon Naturally Speaking 8 (DNS8), French edition

# Data Collection Protocol (4)

## Preparatory tasks:

- Measure base typing speed.
- Standard DNS8 audio enrolment (avg. 6 mins audio per subject).
- 15 minutes demo on how to dictate translations (included vendor recommended error-correction procedure).
- 15 minutes practice session, dictating translation of a similar document to ST1 and ST2.
- Read ST1 and ST2, and carry out terminology and phraseology searches (max 30 mins). This is a standard best practice for translation dictation.
- **Domain Adaptation using vendor provided facilities:**
  - 3.9 million words of Canadian Hansard, for days immediately preceding (but not including) the day for ST1 and ST2

## **Data Collection Protocol (5)**

Data collected:

- 49 mins enrolment French speech
- 81 mins French speech from the ASR task, once long pauses were removed (5478 words)
- screen capture for both ASR and Keyboard tasks.
- audio from debriefing interview

# Productivity measure

Not Word Error Rate!!! (but we still measured it)

Translated Words per Minute (TWPM) is what translators care about:

$$\text{TWPM} = W / T$$

$$T = D + C$$

- $W$  = number of source text words translated
- $T$  = time taken for translation
- $D$  = time spent dictating (for ASR task only, includes time spent thinking)
- $C$  = time spent correcting ASR errors
- Separation between  $D$  and  $C$  done by viewing screen capture.

## Productivity measure (2)

Interpolation used to compute TWPM' of ASR with any given WER'.

Assume correction time of each subject is proportional to WER:

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

Indicative model only:

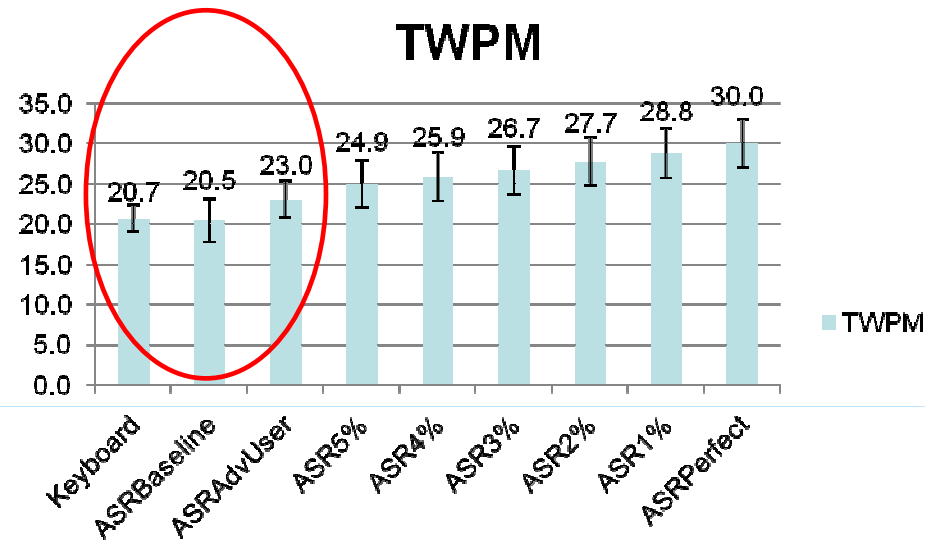
- Assumes relative improvements are the same for all subjects.
- Assumes that all type of errors will improve at same rate.





**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?

# Productivity of ASR vs Keyboard

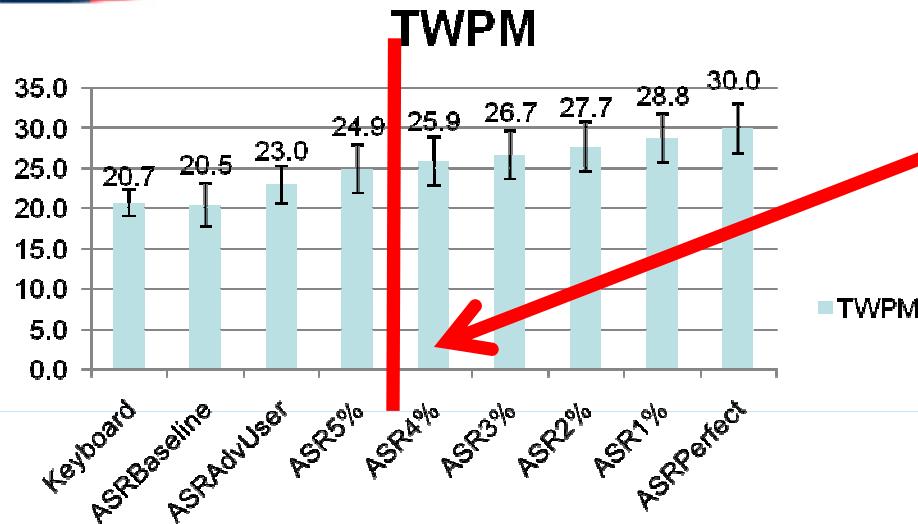


Baseline WER = 11.7%.

ASR no better than keyboard, even when compensating for subjects' lack of familiarity with error-correction procedure (*ASRAdvUser*).

**ASR productivity lower on average**, and for 5 of the 8 subjects in particular.

## Productivity of ASR vs Keyboard (2)



Statistically significant  
breakeven point.

ASRs with WER  $\leq 4\%$  would have exhibited statistically significant productivity gains.

Order of 25.1% to 44.9% relative gains.

## **Productivity of ASR vs Keyboard (3)**

**Sheds serious doubts on the very large two or threefold productivity increases claimed anecdotally.**

Even the most conservative estimates put relative gains at 44.9% TWPM at the maximum.

The three subjects for whom ASR was better, exhibited improvements of 34.8%, 37.8% and 17.6%, which still falls way short of those claims.

This included one subject who had been using ASR in her work for many years.

## Productivity of ASR vs Keyboard (4)

At a typical 1% WER, current off-the-shelf English ASR systems would be *39% faster for writing first draft of a translation.*

But text entry only about half the work, which also includes:

- Pre-reading
- Terminology and phraseology searches
- Self-revision
- Gain for the whole end-to-end translation task might be closer to 20%.

## Productivity of ASR vs Keyboard (5)

Numbers may not tell the whole story

- *2 subjects used to dictation (ASR or dictaphone) more productive with ASR than keyboard.*
- Debriefing interviews indicate that subjects liked the technology and would consider using it.
- Even though they had pretty realistic evaluation of the productivity gain they experienced (or lack thereof).
- Faith that things will improve with time?
- Will this faith remain after short honeymoon period?



**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?

# ASR+SMT combinations

Different ASR variants evaluated

*ASRBaseline*: Domain Adaptation (DA) based on 3.9 million words of Hansard, using the vendor provided facilities (normally used to adapt based on user's email and word-processing documents).

*NoDA*: No Domain Adaptation performed.

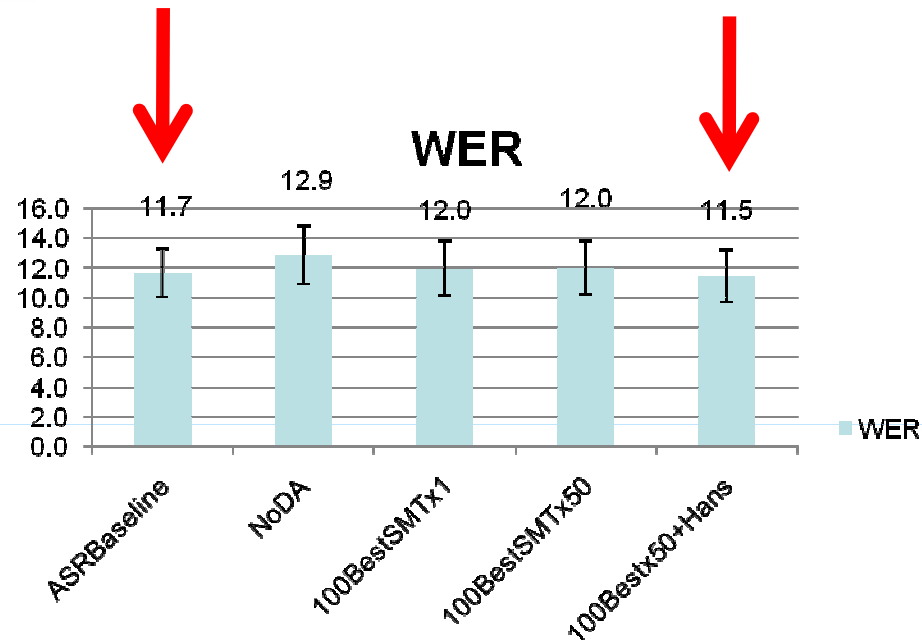
*100BestSMTx1*: DA based on 100 best translations proposed by SMT system (PORTAGE, NRC), for each of the source sentences. No DA based on Hansard.

*100BestSMTx50*: Same as above, except used 50 copies of the 100 best translations, to simulate weighted Language Model Interpolation (LMI).

*100Bestx50+Hans*: Same as above, plus DA based on Hansard.



## ASR+SMT combinations (2)

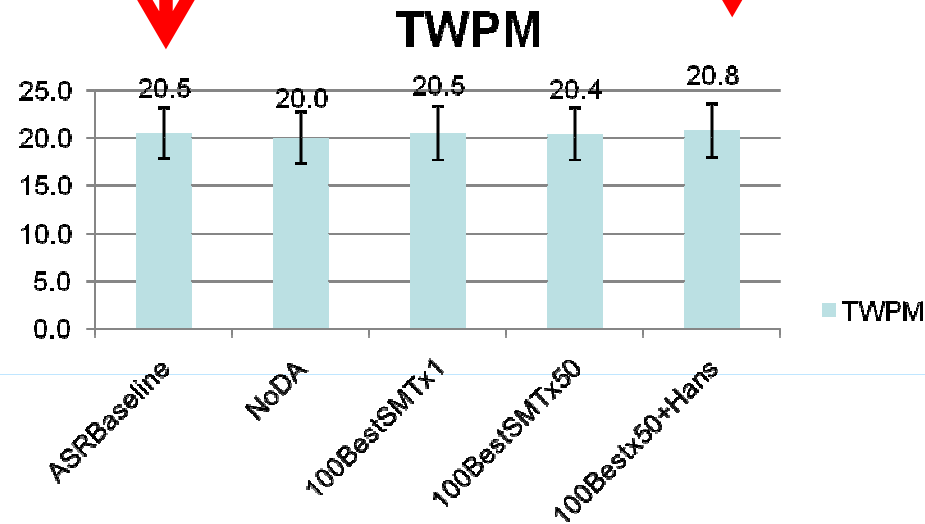


DA based on SMT outputs only had no effect!

Only ASR variants that included Hansard DA were found to be significantly better than *NoDA* variant.

Also, effects of Hansard and MT DA not cumulative

## ASR+SMT combinations (3)



Similar trend if looking at TWPM

Conclusion: ASR+SMT combinations using the **very limited vendor provided** Domain Adaptation facilities does not improve WER nor productivity.

## ASR+SMT combinations (4)

Could tighter ASR+SMT integration make a difference?

Maybe...

- Our maximum relative improvement of 11.7% is much smaller than what was reported by Reddy et al (avg 20.8%) for tighter integration, over a similar benchmark.
- But even a 20.8% relative improvement in WER **would not have resulted** in a statistically significant increase in productivity
- More research needs to be done to actually evaluate tighter integration scenarios.

## Conclusions and Future Research

Very large, two to three-fold productivity gains reported anecdotally, are probably greatly exaggerated.

At WER=11.7%, we found Current French off-the-shelf ASR to not provide statistically significant productivity gains for subjects overall.

But users with dictation experience might still get gains with current technology as is.

Users for whom WER  $\leq$  4% should experience a productivity gain in the order of 25-45% for entering text of a first draft.

Users of recent English ASR in particular, can expect gains in the order of 39%.

## **Conclusions and Future Research (2)**

Limited vendor-provided Domain Adaptation facilities insufficient to improve WER or productivity through ASR+SMT combinations.

More research is needed to evaluate impact of tighter integration on productivity.

**NRC-CMRC**

*Institute for  
Information  
Technology*

# Questions?

