Assessing the Performance of Automatic Speech Recognition Systems When Used by Native and Non-Native Speakers of Three Major Languages in Dictation Workflows

Julián Zapata
School of Translation and Interpretation
University of Ottawa, Canada
jzapa026@uottawa.ca

Andreas Søeborg Kirkedal
Copenhagen Business School
& Mirsk Digital ApS
Denmark
andreas@mirsk.com

Abstract

In this paper, we report on a two-part experiment aiming to assess and compare the performance of two types of automatic speech recognition (ASR) systems on two different computational platforms when used to augment dictation workflows. The experiment was performed with a sample of speakers of three major languages and with different linguistic profiles: non-native English speakers; non-native French speakers; and native Spanish speakers. The main objective of this experiment is to examine ASR performance in translation dictation (TD) and medical dictation (MD) workflows without manual transcription vs. with transcription. We discuss the advantages and drawbacks of a particular ASR approach in different computational platforms when used by various speakers of a given language, who may have different accents and levels of proficiency in that language, and who may have different levels of competence and experience dictating large volumes of text, and with ASR technology. Lastly, we enumerate several areas for future research.

1 Introduction

Speech has been a popular input mode for several years in a number of domains and applications, from automated telephone customer services to legal and clinical documentation. Today, the general problem of automatic recognition of speech by any speaker in any environment is still far from being solved. Nevertheless, speech-enabled interfaces are proven to be more effective than keyboard-and-mouse interfaces for tasks for which full natural language communication is useful or for which keyboard and mouse are not appropriate (Jurafsky and Martin, 2009). Now, although it was implicit in the earliest efforts in natural language processing (NLP) that speech was expected to completely replace — rather than enhance — other input modes, it was soon proposed that, for many tasks, speech input achieved better performance in combination with other input modes (Pausch and Leatherby, 1991).

Clinical documentation and professional translation are two domains in which large volumes of texts are produced on a daily basis worldwide. We carried out an experiment to assess the performance of ASR-augmented dictation workflows using two different computational platforms: a speaker-adapted (SA) PC-based system on a Windows laptop, and a speaker-independent (SI) cloud-based system on an Android tablet. The experimental results of this study may also inform further developments in other areas such as respeaking and live subtitling, where interest in ASR technology has increased in recent years (Romero-Fresco, 2011). The experiment was performed with a small sample of speakers of three different languages and with different linguistic profiles: non-native English (Indian-accented and Spanish-accented) speakers; non-native French (Russian-accented and Spanish-accented) speakers; and native Spanish (Iberian-accented and Latin-American-

accented) speakers. The main objective of this experiment was to examine the potential advantages and drawbacks of ASR-augmentation in different computational platforms and for various users, who may have different accents and levels of proficiency in their working languages, and who may have different levels of competence and experience dictating large volumes of text, and with ASR technology. The general conclusion is that, although some technical challenges still need to be overcome, speech-enabled interfaces have the potential to become one of the most efficient and ergonomic environments to perform translation and documentation tasks (including information retrieval) for an array of users, in addition to other emerging input modes such as gaze, touch and stylus, which may also be combined with speech in multimodal environments (Oviatt, 2012; Zapata, 2014). Lastly, this paper enumerates several areas for future research.

2 Dictation background

As mentioned above, clinical documentation and translation are two domains in which large volumes of texts are produced on a daily basis, and constitute the focus for the present paper. In this section, we provide some background on the use of medical dictation (MD) and translation dictation (TD).

2.1 Medical dictation

A clinical documentation workflow has the following steps: Patient consultation, diagnosis, dictation of diagnosis (using a recording device), transcription and documentation, as illustrated below:

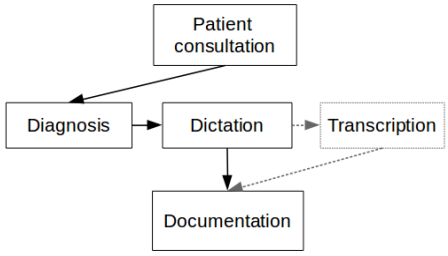


Figure 1. Clinical documentation workflow

The patient is involved during consultation; the physician is involved during all steps except transcription, which is handled by a specially-trained secretary or transcriptionist. The attending physician should approve the transcription before the documentation step, which the secretary also handles. However, this rarely happens in practice since the transcription is not immediate and the physician will have attended other patients in the meantime. In recent years, most hospitals have moved from paper-based clinical records to electronic medical records (EMR) systems where all documentation is stored. When stored in electronic format, information about patient history, medication, etc., and can be immediately shared with other hospitals in case of emergencies. The actual transcription of dictations is commonly handled on a computer using mouse and keyboard.

2.2 Translation dictation

In a professional translation setting, the scenario is similar. In TD, a translator or a team of translators work in collaboration with a transcriptionist or team of transcriptionists. The translator sight-translates a text and records it into a voice recorder. The recording is then sent (via email or a common server) to a transcriptionist, who transcribes the text as instructed by the translator (the latter also dictates punctuation marks and formatting instructions, etc.). It is the translator who makes the final revision to the text manually. Major modifications or additions to the text, if necessary, are dictated and sent again to the secretary for transcription. Figure 2 illustrates the TD workflow:

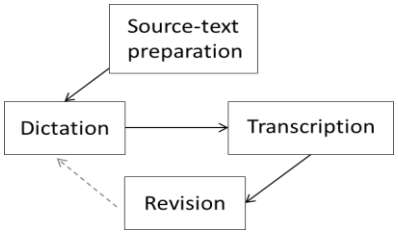


Figure 2. Translation dictation workflow

TD was a very popular – and effective – technique in the 1960s and 1970s (Gingold, 1978), but started to fade away as professional translators' workstations experienced the massive influx of typewriters and personal computers: it was no longer necessary to train and pay additional staff to transcribe translated texts; translators were now able to carry out the transcription by themselves. This being said, a few translation services still opt for this technique in an effort to provide translators with more ergonomic solutions and to increase productivity (Gouadec, 2007; Hétu, 2012). Today, the tremendous improvements in ASR

technology provide a golden opportunity to bring back dictation to the profession; in the words of Gouadec (2007), TD "will become the norm once again".

In MD and TD, the transcription step is slow and expensive. For instance, in MD, the transcription can take between hours and months to complete. Training secretaries to transcribe dictations is expensive and transcription takes up to 60% of secretaries' working hours. Likewise, in TD, it has become difficult to find skilled personnel to type large volumes of texts in a way that the translator-transcriptionist collaboration is cost-effective. Because the transcription will be automatized and immediate with ASR-augmentation, physicians and translators will have the time and the possibility to proofread and approve the transcriptions. In the next section, we provide a brief historical overview of the interest in ASR for TD, and an overview of the different types of ASR systems and of their functioning, while supporting the idea of efficiently integrating this technology to MD and TD workflows.

3 Related work

The interest in ASR technologies for dictation in fields such as translation is not new. Off-the-shelf ASR systems have been part of certain translators' toolbox for over a dozen years now (Bowker, 2002); in many cases, of those translators who once dictated with the aid of voice recorders and transcriptionists back in the 1960s and 1970s.

In the mid-1990s, research efforts to adapt ASR technology to human translation took place for the first time. Such developments focused on minimizing word error rates by combining ASR and machine translation (MT). Hybrid ASR/MT systems have access to the source text and use MT probabilistic models to improve recognition. A number of works have been conducted over the years, highlighting the various challenges of ASR/MT integration (Brousseau et al., 1995; Désilets et al., 2008; Dymetman et al., 1994; Reddy and Rose, 2010; Rodriguez et al., 2012; Vidal et al., 2006), and the potential benefits of using speech input for human translation and post-editing purposes (Garcia-Martinez et al., 2014; Mesa-Lao, 2014). Likewise, further efforts have been made by translation trainers and researchers to evaluate the performance of students and professionals when using off-the-shelf ASR systems for straight TD (Dragsted et al., 2009; Dragsted et al., 2011; Mees et al., 2013); and to assess and analyze professional translators' needs and opinions vis-à-vis ASR technology (Ciobanu, 2014; Zapata, 2012). But ASR systems are not all created equal, and it becomes necessary to investigate what type of system and what conditions of use are more appropriate for the needs of various users in a given domain.

There are three different types of ASR systems wrt. speakers: SI, speaker-dependent (SD) and SA. SI systems use data from many speakers across age, gender, sociolect and dialect to train acoustic models, as well as speaker normalization techniques such as Cepstral Mean and Variance Normalization, Vocal Tract Length Normalization and Maximum Likelihood Linear Transforms (see e.g. Uebel et al. (1999)). Normally, the speaker(s) who will use the system is not in the training data. SD systems are equivalent to SI systems, but use only training data from a single speaker who will also be the sole user of the system. This will produce better recognition performance than SI systems, but the drawback of SD systems is that the amount of training data necessary to train acoustic models is usually not available and time-consuming to collect.

SA systems constitute a middle road. The idea is to adapt an SI system to a specific user using only a little speaker-specific data. Speaker Adaptive Training (SAT) techniques such as Constrained Maximum Likelihood Linear Regression modify either the ASR model parameters or transform the training data directly. See Woodland (2001) for a review of adaptation techniques. Speaker-adaptation of an SI system practically happens in a supervised fashion where the user reads aloud a number of sentences. In this manner, the adaptation software has a gold standard to compare to ASR output and is able to learn a mapping function that optimizes ASR accuracy.

Training ASR systems is a computationally expensive process and training commercial systems can only take place on servers or clusters. Still, the training process can take days. Trained models can be embedded in a system on a computer or can be used from a server accessed through the cloud. The trade-off between embedded vs. cloud is one of computation vs.

latency/connectivity. The computation required in speech decoding (actual recognition) is also expensive. This is not a problem for computers connected directly to a power source, but for laptops and tablets, decoding drains the battery and consumes memory to such an extent that ASR is not a practical tool. Work on reducing memory usage and still achieve acceptable ASR accuracy has been conducted (e.g. in Lei et al. (2013)), but subjects such as reducing computation and implications for battery life have not been addressed.

If speech is streamed to a server, decoded and the output returned via the cloud or intranet, electrical and computational power is abundant. However, the client computer must have fast web access to stream sound to the server and receive text. Latency in ASR confuses users, who will stop dictating, repeat words, restart or speak slower than their natural rate of speech. The problem of battery lifetime can be alleviated by professional translators or physicians who use a desktop computer for dictation and manual revision. This chains the user to the workstation and is not appropriate for cases where mobility is desirable, e.g., physicians who will often have to dictate medical diagnoses while moving from one patient to the next; or translators who need to find ergonomic alternatives to prevent mental and physical fatigue, or even short- and long-term illnesses such as back pain or repetitive stress injury.

4 Research question

The two-part experiment was carried out particularly with translation and medical settings in mind, currently characterized by the extended use of keyboard-and-mouse graphical user interfaces. The underlying hypothesis is that speech input provides one of the most efficient means to perform TD and MD tasks, since ASR "has the potential to be a better interface than the keyboard for tasks for which full natural language communication is useful or for which keyboards are not appropriate" (Jurafsky and Martin, 2009).

But is ASR always beneficial for any task, for any user, in any environment and in any computational platform available today? This is the question that motivates this exploratory study, and is partially answered in the present paper.

5 Experimental setup

This experiment included a sample of English, French and Spanish speakers, as described below:

- English: four Indian-accented speakers (EN1, EN2, EN3, EN4) and one Spanish-accented speaker (EN5) (all non-native)
- French: one Spanish-accented speaker (FR1) and two Russian-accented speakers (FR2, FR3) (all non-native)
- Spanish: two Iberian-accented speakers (SP2, SP3) and two Latin-American-accented speakers (SP1, SP4) (all native)

All participants possess an excellent professional command of the experimental language, whether they speak it as a first, second, third or fourth language. The 12 participants (all graduate students or researchers), had in common at least a minimum level of familiarity with the notions of translation processes, computational linguistics and NLP. However, only a few reported they had hands-on experience with commercial or research-level ASR systems (and were therefore familiar with voice commands, etc.).

5.1 Methodology

Four tasks were involved in the main experiment: (1) typing; (2) reading aloud; (3) dictating with a commercial PC-based SA ASR system[1] on a laptop; and (4) dictating with a commercial cloud-based SI ASR system[2] on a tablet. Tasks 1 and 2 were control tasks, whereas tasks 3 and 4 were the experimental tasks[3].

A 200-word text was chosen for each language. The same text was used for all four tasks. The texts were selected (and amended) so that they would contain the same number of words, one title, two paragraphs and no foreign-language tokens that may not be recognized by an ASR system. For instance, in the English text, a foreign-language name was replaced by "John

[1] Dragon NaturallySpeaking Premium Edition, v.12.5., by Nuance Communications.

[2] Dragon Dictate, integrated in the Swype keyboard, by Nuance Communications.

[3] It was only possible to perform all 4 tasks with the English and French participants, since a PC-based ASR system in Spanish was not available for this experiment.

Smith" so that it would be easily recognized by the ASR systems. In addition, the three texts were relatively simple and contained no specialized terminology. They all contained a fair number of punctuation marks, which needed to be dictated (e.g. "full stop", "comma", "ellipsis", "open quote", "end of quote", "colon") during the experimental tasks (in addition to "new paragraph"). Furthermore, Translog II was used to display the 200-word text in the four tasks and to log the typing session (task 1). Although Translog II was primarily designed to investigate human translation processes, it can also be used to study reading and writing processes in general (Carl, 2012), as in the case of this experiment. This being said, the focus of this experiment was not on keystroke activity but rather on task times and ASR performance across various users, languages and devices.

The main experiment took place over two days. The experimental sessions were performed individually (i.e., one participant at a time). Control tasks were performed separately from experimental tasks (i.e., on different days). This would avoid mental and physical fatigue since each task involved the same text (i.e., typing it, reading it, dictating it on the laptop and dictating it on the tablet). Each task was timed using a stopwatch. No recording of the reading task (task 2) was made. As far as the experimental tasks are concerned (tasks 3 and 4), the transcriptions by the ASR systems both on the laptop and the tablet were saved as Word (.doc) documents.

To measure the word accuracy for the ASR systems, a simple online edit distance calculator[4] (aka. Levenshtein edit distance (Navarro, 2001) calculator) was used. Such a tool calculates the "cheapest" way to transform one string into another. The result obtained indicates the "total cost" or, in other words, the minimum total number of keystrokes what would be needed to edit a given text (in the case of our experiment, the output of the ASR system) to match another text (in our case, the original text).

Lastly, at a later date, a second experimental session took place with the participation of three informants (one per language) from the main experiment[5]. This time, participants were required to proofread and post-edit, using a laptop's keyboard, the texts they had produced earlier with the two ASR systems; in other words, to manually fix the ASR errors. This task was logged using InputLog, a research-level program designed to log, analyze and visualize writing processes (Leijten and Van Waes, 2013). InputLog analyses provide data such as total time spent in the document (i.e. reading through the text and manually fixing the ASR errors), total time of actual keystrokes (additions, deletions and substitutions), total characters typed, switches between mouse and keyboard, etc.

An overview of the results of the experiment and a discussion are provided in the following sections.

6 Results

6.1 Task times and accuracy

It is not surprising that speaking is faster than typing (Hauptmann and Rudnicky, 1990). Our data shows that participants, regardless of whether they were performing the experiment in their native language or in a foreign language, are consistently slower when typing than when reading out loud only. This being said, the reading times across participants and across languages are comparable with mean reading time of 84.74s and standard deviation (SD) of 0.78s. In other words, as it can be observed in Figures 3 and 4 below, it takes about the same amount of time to read a 200-word text in English, in French or in Spanish, whereas typing the same text can take 3-7 times longer.

Figures 3 and 4 also feature task times for ASR tasks. With the exception of EN3, task times for both ASR tasks are comparable (since essentially they involved doing exactly the same thing). ASR task times have longer duration than the reading task because the user needs to dictate punctuation marks and other editing commands. The difference between reading and dictation times (SRT) is statistically significant at *p*-value = 0.0022 measured across all participants. This is unsurprising when comparing SRT mean (128.3s) and SD (29.57s) to reading aloud. A Wilcoxon signed rank-test was used to calculate

[4] The tool, developed by Peter Kleiweg, is available for free at: http://odur.let.rug.nl/kleiweg/lev/.

[5] Unfortunately, the other nine participants were no longer available to perform this task.

the *p*-value because normal distribution of task times cannot be assumed and, with a small sample size, a robust method is needed to calculate statistical significance.

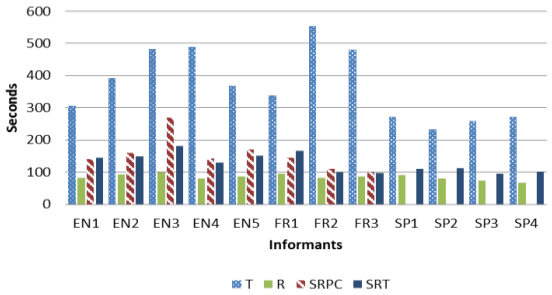


Figure 3. All task times (in seconds). T= typing; R= reading; SRPC: speech recognition on PC; SRT= speech recognition on tablet

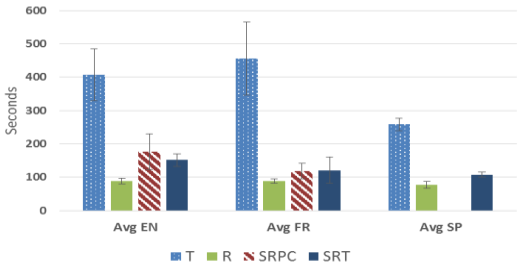


Figure 4. Average task times (in seconds) per language, displaying standard deviation bars.

Tables 1 and 2 show the WAcc for the PC-based system during task 3 in English and French respectively, for each non-native participant (see also Figures 5 and 6 below). It is important to note that the SA system was adapted with minimal training (for approx. 5 minutes) prior to performing the task.

	EN1	EN2	EN3	EN4	EN5
%WAcc-PC	89.2	86.56	80.7	78.97	86.31

Table 1. WAcc on laptop for English language

	FR1	FR2	FR3
%WAcc-PC	95.34	91.76	89.39

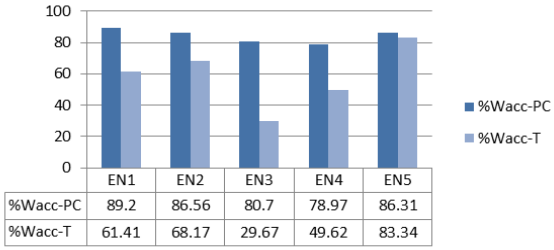
Table 2. WAcc on laptop for French language

For the Spanish language, ASR data was collected with the tablet only. Figure 5 displays very high WAcc rates with the cloud-based SI ASR system ─ with no previous training ─ used by native speakers of the language.

	SP1	SP2	SP3	SP4
%WAcc-T	99.09	97.04	98.85	94.18

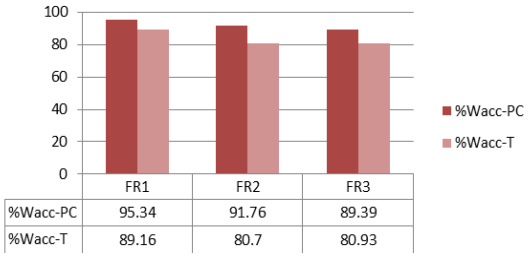
Table 3. WAcc on tablet for Spanish language

Nonetheless, a poor performance of the cloud-based system is observed when used by non-native speakers, particularly Indian-accented English speakers. Figures 5 and 6 display the performance gap between the SA and SI ASR systems and is supported by the difference in means and SD in Table 4.



	EN1	EN2	EN3	EN4	EN5
%Wacc-PC	89.2	86.56	80.7	78.97	86.31
%Wacc-T	61.41	68.17	29.67	49.62	83.34

Figure 5. WAcc on laptop vs. tablet for English language



	FR1	FR2	FR3
%Wacc-PC	95.34	91.76	89.39
%Wacc-T	89.16	80.7	80.93

Figure 6. WAcc on laptop vs. tablet for French language

	EN	SP	FR
Mean (%)	58.44	97.26	83.60
SD (% points)	20.18	2.27	4.82

Table 4. WAcc statistics per language

6.2 Dictation workflow comparison

As mentioned in the methodology section, the focus of the main experiment was to collect data for task completion times and ASR WAcc rates. To compare MD and TD workflows using a transcriptionist to ASR-augmented MD and TD, a revision/post-editing phase must be included in our model of the workflow. ASR, whether SI or SA, is not perfect. We conducted an additional experiment in order to estimate the time and effort that would be required by the user to proof-read and edit the ASR output. This smaller experiment was carried out with informants EN5, FR1 and SP1, who manually post-edited, using a mechanical keyboard, the texts they had previously produced with the SA and SI ASR systems. Figure 7 below shows the time spent typing corrections with the keyboard versus the total time spent proofreading the document. The bars at 100% help illustrate the amount of time spent typing corrections (KT, dark blue) in comparison with the total time spent in the document (full bar).

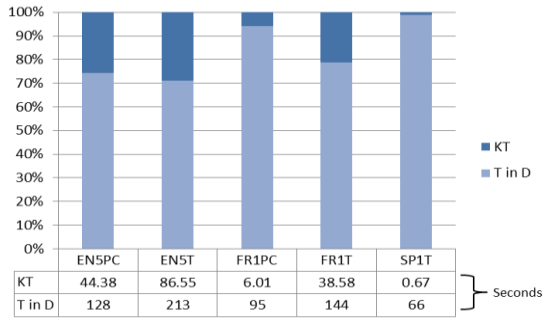


Figure 7. Total time spent typing vs. total revision time (all in seconds). KT= keyboard time; T in D= time in document

Lastly, we added the total revision time to the time dictating with ASR while dictating commands from Figure 3. In short, this indicates the total time a participant would need to carry out a dictation task from start to finish. Thus, as illustrated in Figure 8, this calculation allows us to figure out how much faster it may be to dictate with ASR and manually fix the ASR errors than it is to simply type the text on a mechanical keyboard, and to speculate about possible ways to reduce that time in order to near the reading-out-loud-only times.

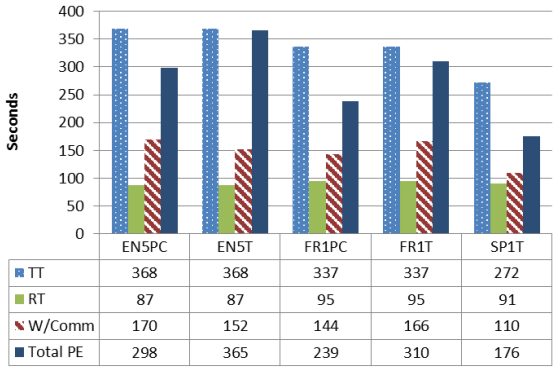


Figure 8. Total task time comparisons for participants EN5, FR1 and SP1 for both ASR systems. TT= typing time; RT= reading time; W/Comm= with commands; Total PE= total time after post-editing

Table 5 provides data on efficiency gains when reading out loud only as compared to typing, and when using ASR and manually post-editing as compared to typing. It also recalls the WAcc for each participant in the different ASR conditions. It can be observed, for instance, that participant EN5, who is a non-native speaker of English, can read out loud an English text 4.22 times faster than she can type the same text. However, with 83.34% WAcc (with the SI system) and 86.31% WAcc (with the SA system), the efficiency gains can be between 1.008 (almost null) and 1.234 respectively.

	T/R	T/Total PE	WAcc (%)
EN5PC	4.22	1.234	86.31
EN5T	4.22	1.008	83.34
FR1PC	3.52	1.41	95.34
FR1T	3.52	1.087	89.16
SP1T	2.99	1.545	99.09

Table 5. Efficiency gains comparison and WAcc for participants EN5, FR1 and SP1. T/R= efficiency gains when reading out loud vs. typing; T/Total PE= efficiency gains after post-editing ASR output vs. typing

In the following section, we provide a discussion on the results of this pilot experiment and formulate areas for future work.

7 Discussion and future work

We have confirmed in this experiment that speaking (or rather, reading aloud) is always faster (approx. 3-7 times) than typing; and we observed that, in terms of efficiency, non-native speakers of a language could benefit from ASR to perform dictation tasks only with an SA system; that native speakers get the best ASR performance (avg. 97% Wacc with the SI system); and that participants who are familiar with ASR technology may benefit considerably from it, regardless of the computational platform they are using (as was the case for EN5, FR1, SP1, SP2 and SP3). In addition, we observed that the extra time to dictate commands (e.g., punctuation marks) is significant and adds to the time needed to post-edit the ASR output. We have also observed that with relatively low WAcc rates (as it was observed for EN5 with the SI system, with 83.3% WAcc) the efficiency gains from ASR-augmentation disappear, but is not less efficient than typing. This being said, to perform certain tasks, punctuation and formatting commands might not always be necessary, or could be avoided using multimodal interaction.

Our experiment models ASR-augmented dictation workflows in two separate stages: a dictation stage and a post-editing stage. This follows the professional translation style taught at most universities: 1) skim the source text, 2) read and comprehend/prepare the source text, 3) create a draft target text, 4) post-edit target text. Our experiment models steps 3 and 4. However, there are many styles of text production and it is highly feasible that a translator or a physician would change errors on-the-fly rather than

complete the dictation first and then proofread and edit the text. Because rereading ASR output to detect errors is unnecessary, post-editing task times (T in D, Figure 9) could be significantly reduced. To test this assumption, an experiment studying on-the-fly editing of dictated text will be conducted. It is a more accurate model of MD and TD workflows without third-party transcription stage and it would be possible to study the pros and cons of multimodal interaction using touch screens, gaze and mouse-and-keyboard. But also comparisons between mechanic keyboards, software keyboards and swipe keyboards will be possible.

The available software and hardware when conducting our experiments were a laptop and a tablet with an SA and an SI ASR system, respectively. The WAcc when using the tablet is consistently lower for non-native speakers of a language, as is expected for an SI system vs. an SA system. Some of the difference can also be due to the different microphones used: for the SI experiments with the tablet, the built-in microphone was used; for the SA experiments with the laptop, a Logitech h600 wireless headset was used. A control experiment with an SI system on the laptop and a SA system on the tablet will follow to shed light on this matter.

With a small number of participants, it is difficult to generalize and draw conclusions based on statistics. To add to our observations, additional experiments with a larger group of informants need to be conducted to make better use of statistical tools and analyses. In addition, larger-scale experiments would need to include longer texts, or several texts following each other, in order to investigate phenomena such as dictation fatigue. Lastly, it would be necessary to include other data collection methods and tools such as video and screen recording, eye-tracking and interviews, and to provide a wider picture of the usability of a particular system or interface by achieving a better understanding and assessment of the correlation between the different aspects of usability (effectiveness, efficiency and user satisfaction), and between objective and subjective usability measures (Hornbæk, 2006).

8 Conclusion

In this experiment, we examined the possibility for native and non-native speakers of a language to use speech as an input modality to dictate large volumes of texts, particularly in clinical documentation and translation workflows. In addition, we were interested in comparing two different ASR environments: a speaker-adapted ASR system installed on a laptop PC and a speaker-independent ASR system in a remote server accessible through a mobile device. We have observed that ASR-augmentation may not be counter-productive. For native speakers, speaker-adaptation does not seem to be necessary to realize efficiency gains, while it is appropriate for non-native speakers. According to our experiments, a WAcc above 83.3% is necessary for the ASR-augmented dictation workflow to be more efficient than typing. Furthermore, we have seen that small differences in WAcc can have a large impact on efficiency. Lastly, we acknowledge that the removal of out-of-vocabulary (OOV) words from the original English text (i.e. replacing a foreign name with "John Smith") may have biased the results in favour of the ASR solution because OOV words can have a large impact on WAcc. In real-life translation and medical dictation tasks with many proper names, pharmaceuticals and new terms, OOV words are more likely to occur frequently and that failed recognition of OOVs or recognition of different words can have a large impact on WAcc.

On one hand, as hospitals continue moving towards EMR systems and more efficient clinical documentation becomes necessary; and, on the other hand, as web-based translation tools and environments become more and more popular and efficient; it becomes essential to closely examine the different text-input modalities available in keyboard-less devices, as we move towards the era of mobile computing and ubiquitous information.

Acknowledgments

A special thanks to all anonymous participants for their time. We also acknowledge the assistance of Maheshwar Ghankot and guidance provided by Srinivas Bangalore and Michael Carl during the first set of experiments, and for their comments on an earlier version of this paper. This study was carried out within the framework of the translation data analytics project held in July-August 2014 at the Copenhagen Business School, in Denmark. The project was supported by the European Union's 7th Framework Program (FP7/2007-2013) under grant agreement 287576 (CASMACAT).

References

Bowker, Lynne. 2002. *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

Brousseau, Julie, Caroline Drouin, George Foster, Pierre Isabelle, Roland Kuhn, Yves Normandin, and Pierre Plamondon. 1995. "French Speech Recognition in an Automatic Dictation System for Translators: The TransTalk Project." In *Proceedings of Eurospeech'95*. http://www.iro.umontreal.ca/~foster/papers/ttalk-eurospeech95.pdf.

Carl, Michael. 2012. "Translog - II: A Program for Recording User Activity Data for Empirical Reading and Writing Research." In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, 4108–4112.

Ciobanu, Dragoș. 2014. "Of Dragons and Speech Recognition Wizards and Apprentices." *Revista Tradumàtica* (12): 524–538.

Désilets, Alain, Marta Stojanovic, Jean-François Lapointe, Rick Rose, and Aarthi Reddy. 2008. "Evaluating Productivity Gains of Hybrid ASR-MT Systems for Translation Dictation." In *Proceedings of the IWSLT2008*. http://www.mt-archive.info/IWSLT-2008-Desilets.pdf.

Dragsted, Barbara, Inge Gorm Hansen, and Henrik Selsøe Sørensen. 2009. "Experts Exposed." *Copenhagen Studies in Language* 38: 293–317.

Dragsted, Barbara, Inger M. Mees, and Inge Gorm Hansen. 2011. "Speaking Your Translation: Students' First Encounter with Speech Recognition Technology." *Translation & Interpreting* 3 (1): 10–43. http://www.trans-int.org/index.php/transint/article/viewFile/115/87.

Dymetman, Marc, Julie Brousseau, George Foster, Pierre Isabelle, Yves Normandin, and Pierre Plamondon. 1994. "Towards an Automatic Dictation System for Translators: The TransTalk Project." In *Fourth European Conference on Speech Communication and Technology*, 4. Citeseer. http://arxiv.org/abs/cmp-lg/9409012.

Garcia-Martinez, Mercedes, Karan Singla, Aniruddha Tammewar, Bartolomé Mesa-Lao, Ankita Thakur, M. A. Anusuya, Michael Carl, and Srinivas Bangalore. 2014. "SEECAT: ASR & Eye-Tracking Enabled Computer-Assisted Translation." In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, 81–88.

Gingold, Kurt. 1978. "The Use of Dictation Equipment in Translation." In *La traduction, une profession. Actes du VIIIe Congrès mondial de la fédération internationale des traducteurs*, edited by Paul A. Horguelin, 444–448. Ottawa: Conseil des traducteurs et interprètes du Canada.

Gouadec, Daniel. 2007. *Translation as a Profession*. Amsterdam: John Benjamins.

Hauptmann, Alexander G., and Alexander I. Rudnicky. 1990. "A Comparison of Speech and Typed Input." In *Proceedings of the Speech and Natural Language Workshop*, 219–224.

Hétu, Marie-Pierre. 2012. "Le travail au dictaphone, une solution ergonomique?" *Circuit* 116 (summer 2012): 23.

Hornbæk, Kasper. 2006. "Current Practice in Measuring Usability: Challenges to Usability Studies and Research." *International Journal of Human-Computer Studies* 64 (2) (February): 79–102. doi:10.1016/j.ijhcs.2005.06.002. http://linkinghub.elsevier.com/retrieve/pii/S1071581905001138.

Jurafsky, Daniel, and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Lei, Xin, Andrew Senior, Alexander Gruenstein, and Jeffrey Sorensen. 2013. "Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices." *Interspeech* (August): 662–665. http://research.google.com/pubs/archive/41176.pdf.

Leijten, Mariëlle, and Luuk Van Waes. 2013. "Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes." *Written Communication* 30 (3) (June 29): 358–392. doi:10.1177/0741088313491692. http://wcx.sagepub.com/cgi/doi/10.1177/0741088313491692.

Mees, Inger M., Barbara Dragsted, Inge Gorm Hansen, and Arnt Lykke Jakobsen. 2013. "Sound Effects in Translation." *Target* 25 (1) (January 1): 140–154. http://openurl.ingenta.com/content/xref?genre=article&issn=0924-1884&volume=25&issue=1&spage=140.

Mesa-Lao, Bartolomé. 2014. "Speech-Enabled Computer-Aided Translation: A Satisfaction Survey with Post-Editor Trainees." In *Workshop on Humans and Computer-Assisted Translation*, 99–103.

Navarro, Gonzalo. 2001. "A guided tour to approximate string matching". *ACM Computing Surveys*, 33(1): 31-88.

Oviatt, Sharon. 2012. "Multimodal Interfaces." In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, edited by Julie A. Jacko,

3rd ed., 415–429. New York: Lawrence Erlbaum Associates.

Pausch, Randy, and James H. Leatherby. 1991. “An Empirical Study: Adding Voice Input to a Graphical Editor.” *Journal of the American Voice Input/Output Society* 9 (2): 55–66.

Reddy, Aarthi, and Richard C. Rose. 2010. “Integration of Statistical Models for Dictation of Document Translations in a Machine Aided Human Translation Task.” *IEEE Transactions on Audio, Speech and Language Processing* 18 (8): 1–11. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05393062.

Rodriguez, Luis, Aarthi Reddy, and Richard Rose. 2012. “Efficient Integration of Translation and Speech Models in Dictation Based Machine Aided Human Translation.” In *Proceedings of the IEEE 2012 International Conference on Acoustics, Speech, and Signal Processing*, 2:4949–4952.

Romero-Fresco, Pablo. 2011. *Subtitling Through Speech Recognition: Respeaking*. Manchester: St. Jerome.

Uebel, Luis Felipe, and Philip C. Woodland. 1999. “An Investigation into Vocal Tract Length Normalisation.” In *Sixth European Conference on Speech Communication and Technology*, 1–4. http://www.isca-speech.org/archive/eurospeech_1999/e99_2527.html.

Vidal, Enrique, Francisco Casacuberta, Luis Rodríguez, Jorge Civera, and Carlos D. Martínez Hinarejos. 2006. “Computer-Assisted Translation Using Speech Recognition.” *IEEE Transactions on Audio, Speech and Language Processing* 14 (3). http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01621206.

Woodland, Philip C. 2001. “Speaker Adaptation for Continuous Density HMMs: A Review.” In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*.

Zapata, Julián. 2012. “Traduction dictée interactive : intégrer la reconnaissance vocale à l’enseignement et à la pratique de la traduction professionnelle.” M.A. thesis, University of Ottawa. http://www.ruor.uottawa.ca/en/bitstream/handle/10393/23227/Zapata Rojas_Julian_2012_these.pdf?sequence=1.

Zapata, Julián, 2014. “Exploring Multimodality for Translator-Computer Interaction.” In *Proceedings of the 16th International Conference on Multimodal Interaction*, 339–343. http://dl.acm.org/citation.cfm?id=2666280.