The Use of Machine-Generated Transcripts During Human Translation

Allison L. Powell
Corporation for National
Research Initiatives
Reston, VA, USA
apowell@cnri.reston.va.us

Allison Blodgett
University of Maryland
Center for Advanced Study of Language
College Park, MD, USA
ablodgett@casl.umd.edu

Abstract

At the request of the USG National Virtual Translation Center, the University of Maryland Center for Advanced Study of Language conducted a study that assessed the role of several factors mediating transcript usefulness during translation tasks. These factors included source language (Mandarin or Modern Standard Arabic), native speaker status of the translators, transcript quality (low or moderate word error rate), and transcript functionality (static or dynamic). Using 54 Mandarin and 54 Arabic translators (half native speakers in each language) and broadcast news clips for input, the study demonstrated that translation environments that provide dynamic transcripts with low or moderate word error rates are likely to improve performance (measured as integrated speed and accuracy scores) among non-native speakers without decreasing performance among native speakers.

1 Introduction

One goal of human language technology is to improve human performance on real world tasks. In meeting this goal, it is useful to know which aspects of technology improve performance, whether those aspects affect some users more than others, and how users respond to features of the technology in general. The current study focuses on machine-generated transcripts because they have the potential to improve how quickly and/or accurately humans translate audiovisual material. For example, translators might have more time to devote to more challenging portions of the material, and thus produce a more faithful written translation, if the transcript enables them to spend less time on easier portions and/or to replay more difficult portions via clicking on characters in a dynamic transcript.

Of course, several factors influence how likely it is for a transcript to support human translators, including—but not limited to—the orthography of the source language, the translator's status as a native or non-native speaker, the transcript's word error rate (WER), and the extent to which the transcript provides capabilities beyond a written record of the input. In this study, we investigate whether providing human translators with machine-generated transcripts improves performance. We manipulate several factors including native speaker status, source language (Mandarin or Modern Standard Arabic), transcript WER (low or moderate), and transcript functionality (static or dynamic).

Translators work within three versions of an integrated translation environment. The simplest version provides audiovisual material (i.e., a broadcast news clip), playback controls, and a place to type the translation. A second version adds a machine-generated transcript with color-coding for important elements of information, such as person, place, and organization names. A final version adds a moving cursor within the transcript, which is synchronized to the audio. In addition to the moving cursor, translators can click on words or characters to jump to that portion of the news clip.

These three translation environments incorporate aspects of Translator's Aide, a translation support tool developed by BBN Technologies and funded by the Department of Defense Technical Support Working Group (TSWG). Translator's Aide is a companion to the BBN Broadcast Monitoring System, which includes an information retrieval component. The Broadcast Monitoring

System is also supported by TSWG and deployed in various government sites working open source analysis. Analysts use this system to quickly find video clips on relevant topics by searching a machine transcript or machine translation. Users can then capture news clips and export them to Translator's Aide for human translation. This study focuses on the use of transcripts in human translation. With TSWG support, BBN created the three translation environments in this experiment.

2 Experiments

We conducted separate experiments for Mandarin and Arabic using a 2 (speaker status: native or non-native) x 3 (translation environment: dynamic transcript, static transcript, no transcript) x 2 (transcript WER: low or moderate) design. We counterbalanced translation environment and WER across participants to help control for effects of translator fatigue or practice.

Our main prediction was that access to transcripts, particularly dynamic transcripts, would improve translator performance for native and non-native speakers. This finding would be consistent with Munteanu et al. (2006) in which native speakers of English showed a marginal improvement in quiz scores when they had access to lecture transcripts with a 25% WER.

The transcript format was that produced by the Broadcast Monitoring System for Translator's Aide: punctuation was included, and while speaker turns were not explicitly marked, speaker changes did correspond to paragraph breaks. Based on the work of Jones et al. (2003), this format is expected to have an acceptable readability.

2.1 Participants

For each language, 27 native and 27 non-native speakers participated. Participants self-reported native speaker status and provided information regarding translation experience and language skills in Mandarin or Arabic, and English.

To ensure that participants would be able to perform the translation tasks, eligibility criteria required them to hold a 2+ or equivalent in Mandarin or Arabic listening. This number comes from the Interagency Language Roundtable (ILR) Scale, which provides proficiency scores for reading, listening, writing, and speaking tests on a scale of 0 to 5. A 2+ corresponds to "limited working proficiency, plus."[1]

In order to recruit translators who are representative of those who work with the National Virtual Translation Center (NVTC) and other USG agencies, the NVTC targeted professional translators in the Washington, DC, and Salt Lake City, UT, areas. All but 6 participants (3 native Mandarin speakers, 1 non-native Mandarin speaker, and 2 native Arabic speakers) reported translation experience.

Table 1 summarizes participants' mean ages and their answers to two questions regarding years of experience with translation, summarization, transcription, or interpretation tasks in general and also with respect to Mandarin or Arabic.

	Mean Age (Range)	Mean Years of Experience (Range)	Mean Years of Experience Specific to Mandarin or Arabic (Range)
Mandarin			
Native	40.7 (19-57)	8.43 (0-40)	7.76 (0-40)
Non-native	33.96 (23-62)	7.44 (0-30)	4.29 (0-20)
Arabic			
Native	43.2 (23-75)	8.54 (0-35)	7 (0-35)
Non-native	30.3 (22-49)	4.74 (1-20)	3.17 (0.6-12)

Table 1. Summary of Participant Ages and Experience

2.2 Task

Each participant took part in a two-day individual or group session. On the first day (4 hours), translators learned to use all the features available in the integrated translation environments and then completed 1 to 2 short practice news clips in each of the three versions: the environment that provides a dynamic transcript, the one that provides a static transcript, and the one that provides no transcript.

[1] See http://www.govtilr.org/Skills/ILRscale3.htm for a full list of ILR skill level descriptions. See Herzog (http://www.govtilr.org/Skills/index.htm) for a description of the ILR Scale.

On the second day (8 hours), translators began with a short warm-up clip using the environment of their choice. They then worked on two test clips—one from the low WER group and one from the moderate WER group—in each of the three environments. Translators had 20 minutes for the short (approximately 1 minute) warm-up clip and 50 minutes for each of the longer (approximately 2 minute) test clips.

At the end of each clip, translators completed a short questionnaire to indicate the degree to which they completed the translation, their perception of clip difficulty, their opinion of the transcript quality (if a transcript had been available), and any comments they might have. At the conclusion of the study, they rank ordered their preferred translation environments.

Each news clip came from Al-Jazeera or Chinese Central Television and contained a coherent news story without commercial break.

Translators had no access to on-line dictionaries, but could use whatever hand-held electronic or hard copy aids they chose to bring to testing.

2.3 Transcript Quality Measures

For each pair of test clips that translators used with a particular translation environment, one corresponded to a low WER group and the other, to a moderate WER group. WERs were naturally occurring, and we identified errors in the conventional way (i.e., on the basis of the number of insertions, deletions, or substitutions required to make a machine-generated transcript identical to an accurate human transcript, divided by the number of words or characters contained in that accurate human transcript).

Table 2 summarizes error rates for each group of clips in each language. It also contains an estimated WER for Mandarin in which character error rate (CER) is multiplied by a factor of 1.6, the mean character length for semantic concepts in Chinese (e.g., Gao, Goodman, Li, & Lee, 2002). We estimated WER in Mandarin in an initial attempt to provide transcripts with similar WER ranges in both languages.

We selected the six test clips in each language from a larger group that met our minimal WER requirements (e.g., not exceeding 30%). We realized that clips with similar WERs might have very different error characteristics. In an attempt to increase consistency within clips of the same language and WER group, we also considered human perceptions of the transcripts by asking a native speaker of Mandarin and a fluent non-native speaker of Arabic to listen to each clip and identify errors in the transcript as either serious (ones that would likely give a reader trouble) or minor (ones that a reader would likely gloss over). These speakers returned different patterns of responses.

	Low	Moderate
Mandarin CERs	6.26	10.93
	8.06	11.97
	9.34	13.43
Mandarin Estimated WERs (CERs*1.6)	10.02	17.49
	12.90	19.15
	14.94	21.49
Arabic WERs	18.30	26.84
	19.47	27.44
	23.62	29.93

Table 2. Word and character error rates (WERs, CERs) for each transcript

The native Mandarin speaker judged the majority of transcript errors to be serious because they resulted in text that did not make sense or that had the wrong meaning. Minor errors generally included names of foreigners and punctuation. In contrast, the Arabic speaker judged the majority of transcript errors to be minor errors, and many of these minor errors reflected omissions of the definite article or misspelled proper names.

This contrast might help explain why our best Arabic WER (18.30%) exceeds the upper bound that we initially established for identifying low WER transcripts (10-15%). Transcripts in Arabic might generally have a higher WER than transcripts in Mandarin given that Arabic (but not Mandarin) uses definite articles and that these articles are problematic for the automatic speech recognizer because they are clitics.

To quantify the human observations, we calculated two measures for each transcript: serious error frequency (SEF) and serious error distance (SED). SEF is the number of serious errors divided by the total number of words/characters in the original machine-generated transcript. This takes into account the number of serious errors while adjusting for text length. SED is the mean length in words/characters of each stretch of text that is free from serious errors divided by total number of words/characters. This takes into account the amount of serious-error-free text available to a translator in the original transcript.

There is some question as to whether it would be more appropriate to use the machine transcript or the human-corrected transcript in adjusting for text length in these calculations. Whereas the former takes into account the state of the transcript that a user encounters, the latter takes into account the amount of spoken material the user accesses in assessing whether an error is serious or minor. The results with our materials were similar either way.

We realize that other factors (e.g., how soon the first serious error occurs, proportion and distribution of minor errors, how well translators read in the language, etc.) likely influence perceived or actual transcript usefulness. Indeed, while SEF and SED rates corresponded neatly with Mandarin WERs, the same could not be said for Arabic. (See the tables in Appendix A for SEF and SED rates and other characteristics of the test clips.) Ultimately, as clips rotated across the three translation environments, each low WER clip had a lower proportion of serious errors (SEF) and a higher distance between serious errors (SED) relative to the corresponding moderate WER clip.

2.4 Editing and Scoring of Translations

To assess translation accuracy, two human judges (from a set of 7-10 judges) compared each completed translation against a reference translation, which served as an answer key. Whenever the translation differed in meaning from the reference, the judges edited the translation to make it convey the same meaning as the reference. We trained the judges to follow a set of guidelines that emphasized editing for meaning only (not for style or polish), while using the fewest number of edits possible. This measure derives in part from the HTER (Human Translation Edit Rate) evaluation approach (Snover et al., 2006) developed to measure the quality of machine translations as part of the Defense Advanced Research Projects Agency (DARPA) Global Autonomous Language Exploitation (GALE) initiative.

A member of the research team acted as an adjudicator to review the two edited translations in conjunction with the reference to ensure that the judges followed the guidelines accurately. Using the judges' edits whenever possible and supplementing when necessary, the adjudicator created a final adjudicated translation. We used tercom version 6b[2] to compare the original translation to its adjudicated version and establish a numerical score indicating how similar the documents were. The lower the numerical score, the fewer edits a translation required, and the more accurately the translation conveyed the meaning in the reference.

2.5 Integrated Speed and Accuracy Scores

We collected two measures of performance: time on task and translation accuracy. The translation environment, which automatically logged start and stop times whenever translators opened and closed broadcast news clips, provided time on task.

Translators knew they were being timed, and we encouraged them to work as quickly as possible. However, as a group, they tended to use the full 50 minutes or close to it. Although time on task does not reveal at what point (if any) participants shifted from writing an initial translation to proofreading the text, in adjudicating translations, we noticed that many were incomplete. This suggests that many translators tended to spend the allotted time writing the initial translation. There were no incentives to finish early. In order to test groups of translators and ensure that they took adequate breaks, each clip began at a scheduled time.

	Mandarin		Arabic	
Translation Environment	Native	Non-native	Native	Non-native
Dynamic Transcript	48	49	44	47
Static Transcript	49	50	43	47
No Transcript	48	48	44	47

Table 3. Summary of Mean Time on Task (in Minutes)

The primary indicator of performance is an integrated score that combines time on task, amount of material to translate, and translation accuracy, as in (1):

(1) Integrated Score = Translation Score * (Time Taken/Time Allowed) / (Current Clip Duration / Shortest Clip Duration)

[2] http://www.cs.umd.edu/~snover/tercom/

This calculation provides two adjustments to translation scores in order to credit translators who finished early and to correct for increases in scores that resulted from longer clips having more content to translate than shorter clips. The first adjustment uses the ratio of time taken to time allowed in order to integrate translation scores with time on task as a single value. The result is that performance scores decrease (or improve) as translators finish a translation early. The second adjustment uses the ratio of current clip duration to shortest clip duration to correct for increases in translation scores that result from longer clips having more content to translate than shorter clips. The result is that scores decrease as clip duration increases.

For each language, we submitted the integrated scores to a subject-based 3 (translation environment: dynamic transcript, static transcript, no transcript) x 2 (transcript WER: low or moderate) repeated measures ANOVA. Native speaker status was a between-subjects variable.

2.6 Mandarin Results

Table 4 summarizes the mean integrated speed and accuracy scores by native speaker status, translation environment, and WER for Mandarin translators. **Lower numerical scores indicate better performance than higher scores.**

	Native Mandarin	Non-native Mandarin
Dynamic Transcript		
Low WER	27.8	39.1
Mod WER	25.5	36.0
Static Transcript		
Low WER	31.1	43.2
Mod WER	26.7	38.2
No Transcript		
Low WER	30.6	47.9
Mod WER	25.3	43.4

Table 4. Mean Integrated Mandarin Scores by Translation Environment and WER Group

The ANOVA results demonstrate a main effect of native speaker status (F[1,52]=9.890, p<.01) such that scores are significantly lower for native speakers than non-native. There is also a main effect of WER (F[1,52]=31.302, p<.01) that is present even in the no transcript condition. Surprisingly, the WER effect is in an unexpected direction. Clips associated with moderate WER transcripts elicited *better* integrated scores than clips associated with low WER ranges.

We have no explanation for this effect. We can report that the measures of translators' perceptions of transcript quality and clip difficulty may reveal a clue. In terms of transcript quality, translators reported poorer than expected assessments of some low WER transcripts and higher than expected assessments of some moderate WER transcripts. In terms of clip difficulty, translators reported that two of the clips with low WER transcripts were particularly challenging. These conflicts might serve as a useful starting point for future discussions of transcript quality with native and non-native speakers of Mandarin.

The results further show a main effect of translation environment (F[2,104]=6.857, p<.01) that is qualified by a significant interaction between environment and native speaker status (F[2,104]=4.146, p<.05).

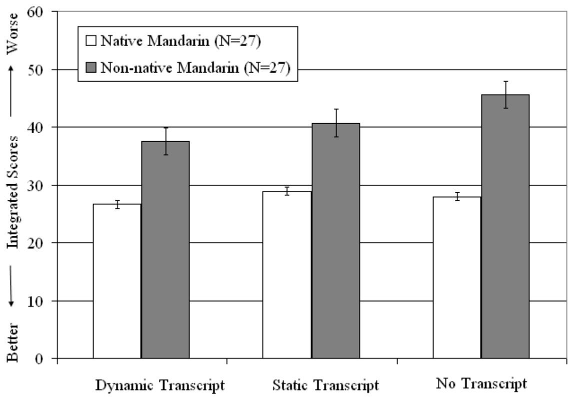

Figure 1. Mean Integrated Mandarin Scores (and Standard Error Bars) by Translation Environment

Because we predicted that transcripts, particularly dynamic ones, would improve performance, we investigated the interaction, which is represented in Figure 1. First, we conducted separate repeated measures ANOVAs for the native speakers (F[2,52]=1.184, p>.10) and non-native speakers (F[2,52]=7.746, p<.01). Then, we investigated the significant effect of translation environment for the non-native speakers through least significant difference pairwise comparisons.[3]

The results of the pairwise comparisons in Table 5 indicate that the dynamic transcripts significantly lower integrated scores compared to the no transcript case. The results further indicate that the gains from introducing a static transcript (i.e., No Transcript to Static) and from further introducing synchronization and navigation features (i.e., Static to Dynamic) are marginally significant.

Comparison	Mean Difference (SE), p value
Dynamic to No Transcript	8.096 (1.961), p<.01
Static to No Transcript	4.938 (2.405), p=.05
Static to Dynamic	3.158 (1.808), p=.09

Table 5. Results of Least Significance Difference (LSD) Pairwise Comparisons and Standard Error (SE) within the Non-native Mandarin Speaker Data

2.7 Arabic Results

Table 6 summarizes mean integrated speed and accuracy scores by native speaker status, translation environment, and WER for Arabic translators.

	Native Arabic	Non-native Arabic
Dynamic Transcript		
Low WER	11.9	21.7
Mod WER	21.7	29.2
Static Transcript		
Low WER	12.0	22.1
Mod WER	16.5	32.2
No Transcript		
Low WER	12.3	24.8
Mod WER	13.8	33.2

Table 6. Mean Integrated Arabic Scores by Translation Environment and WER Group

The ANOVA results demonstrate a main effect of WER (F[1,52]=45.594, p<.01) that was in the expected direction: clips associated with low WER transcripts elicited lower integrated scores than clips associated with moderate WER transcripts. As with Mandarin, we observe the WER effect in the no transcript condition. The results also demonstrate a main effect of native speaker status (F[1,52]=10.621, p<.01) such that native speakers outperform non-natives.

	WER		
	Low	Moderate	
Native Arabic	12.1	15.1	t[81]=-4.021, p<.01
Non-native Arabic	22.9	31.6	t[81]=-7.018, p<.01

Table 7. Mean Integrated Arabic Scores and t-test results by WER and Native Speaker Status

Two interactions qualify these main effects. First, WER interacts with native speaker status (F[1,52]=10.274, p<.01). As shown in Table 7, the results of paired samples two-tailed t-tests suggest that both speaker groups are sensitive to WER. However, the t-values suggest that we can be even more confident that the effect holds among non-native speakers than we can be for native speakers.

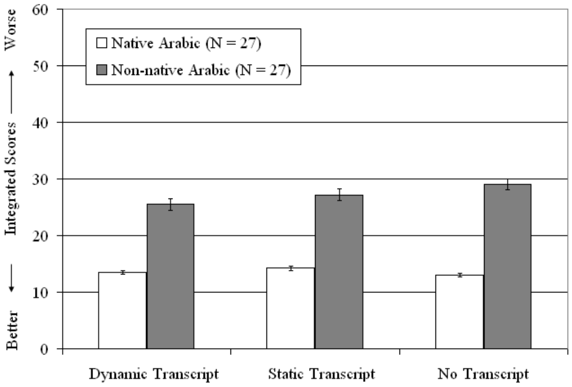

Figure 2. Mean Integrated Arabic Scores (and Standard Error Bars) by Translation Environment

Second, there is a marginal interaction between translation environment and native speaker status (F[2,104]=2.575, p=.08), as shown in Figure 2. Although the interaction is marginal, our main pre-

[3] We adopted this approach in line with Cohen (2001, p. 376), who argues that pairwise comparisons following a significant ANOVA are appropriate for comparisons of three groups. While some might argue that this invites Type 1 error, we would argue that more harm is done in this case by inviting Type II error and incorrectly accepting the null hypothesis that transcripts provide no benefit to this population.

diction is that transcripts, particularly dynamic ones, would improve performance, and we followed the same steps as in the Mandarin analysis.

First, we conducted separate repeated measures ANOVAs for the native speakers (F[2,52]=.769, p>.10) and non-native speakers (F[2,52]=2.518, p=.09). Then, we investigated the marginal effect of translation environment for non-native speakers using Least Significant Difference pairwise comparisons, the results of which are shown in Table 8.

The comparison between the dynamic transcript and the no transcript case approaches significance and is likely the source of the marginal interaction. It is worth nothing that this pattern for non-native speakers is consistent with the significant findings for non-native Mandarin speakers.

Comparison	Mean Difference (SE), p value
Dynamic to No Transcript	3.562 (1.723), p=.05
Static to No Transcript	1.894 (1.560), p=.24
Static to Dynamic	1.668 (1.471), p=.27

Table 8. Results of Least Significance Difference (LSD) Pairwise Comparisons and Standard Error (SE) within the Non-native Arabic Speaker Data

2.8 Translator Preferences

Native and non-native speakers liked the language technology. When asked to rank order their preferences for the three translation environments, 81% of native speakers and 94% of non-native speakers indicated that they most preferred having access to the dynamic transcript. In addition, many translators reported using the low WER and moderate WER transcripts. This widespread use, shown in Table 9, is encouraging given that other clip types (e.g., talk shows, man on the street interviews, etc.) are likely to yield dramatically higher WERs. This study purposely tested transcripts from the high end of the quality spectrum in order to maximize the likelihood of detecting a benefit.

Translators' comments provided further insights. While native and non-native speakers alike commented that they used the transcripts for reference and navigation, non-native speakers added that they used them to get a sense of where a difficult passage was going; to be able to assign the right grammatical relations among parts of long or complex utterances; as an aid to understanding the audio, especially for fast speech; and as an aid in spelling, dictionary look-up, and identification of related words. Multiple translators commented that the transcripts would likely provide an even greater benefit if they could edit the transcript.

	Percent Reporting Transcript Use	
	Low WER	Moderate WER
Mandarin		
Native	93%	93%
Non-native	89%	94%
Arabic		
Native	78%	69%
Non-native	100%	96%

Table 9. Percentage of Translators by Language and Native Speaker Status Who Report Having Used the Transcripts During Translation

A small percentage of translators (6%, evenly divided among native Arabic, native Mandarin, and non-native Mandarin participants) did not find having a transcript useful. More generally, some translators reported not using transcripts because of quality issues, and some commented on the need to be vigilant about noticing when transcripts deviated from the audio. Mandarin translators in particular found errors in punctuation in the transcripts to be especially problematic. Some translators reported simply not needing a transcript (e.g., because a clip was easy for them).

3 Conclusion

Our findings suggest that access to machine-generated transcripts—particularly ones that allow translators to interact dynamically with the audiovisual material—can improve translation performance among non-native speakers. While the benefits seemed to occur even for transcripts with moderate WERs, future work will need to disentangle WER, transcript quality, and clip difficulty in naturally occurring transcripts.

Acknowledgments

This project was a collaborative initiative involving the University of Maryland Center for Advanced Study of Language (CASL), the National Virtual Translation Center (NVTC), the Department of Defense Technical

Support Working Group (TSWG), BBN Technologies, and the Corporation for National Research Initiatives (CNRI).

Dr. Kathleen Egan, as NVTC Director of Technology, Mr. Everette Jordan, as NVTC Director, and the NVTC Operations Team were instrumental in making the task responsive to real USG needs, and they provided the original vision in recognizing the need for a scientific evaluation of the design features intended to support human translators. We are especially grateful to the NVTC, and in particular to Joy Miller and Terra Wathen, as well as Caitlin Christianson and Jon Phillips, for providing participants who were representative of USG settings, as well as various assessments of broadcast materials.

The criteria used to evaluate translator performance were developed in consultation with Dr. Egan and her team, and we further benefited from discussions with researchers at the National Institute for Standards and Technology (NIST). We also acknowledge support from TSWG, who tasked BBN under an existing contract to modify Translator's Aide for use in the study.

We would like to thank the Utah National Guard for the use of their facilities and for support during the experiments conducted there. We are grateful for the participation of the translators and translation judges. Finally, we would like to thank the research and support staff, and in particular, Dr. Amy Weinberg for her technical assistance and leadership.

References

Barry H. Cohen. 2001. *Explaining Psychological Statistics,* 2nd Edition. John Wiley & Sons, Inc., NY, NY.

Jianfeng Gao, Joshua T. Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a Unified Approach to Statistical Language Modeling for Chinese. *ACM Transactions on Asian Language Information Processing*, 1(1):3-33.

Douglas Jones, Florian Wolf, Edward Gibson, Elliott Williams, Evelina Fedorenko, Douglas Reynolds, and Marc Zissman. 2003. Measuring the Readability of Automatic Speech-to-Text Transcripts. *Eurospeech-2003,* 1585-1588.

Martha Herzog, An Overview of the History of the ILR Language Proficiency Skill Level Descriptions and Scale. http://www.govtilr.org/Skills/index.htm. Accessed Jan. 10, 2008.

Cosmin Munteanu, Gerald Penn, Ron Baecker, Elaine Toms, and David James. 2006. Measuring the Acceptable Word Error Rate of Machine-Generated Webcast Transcripts. *Proc Interspeech*, 1-4.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proc Association for Machine Translation in the Americas*.

Appendix A

Tables 10 and 11 provide summaries of the test clip characteristics. The column headers correspond to the following:

(#) Clip Number
(A) Percent Word Error Rate
(B) Number of ASR Characters/Tokens
(C) Number of Serious Errors
(D) Serious Error Frequency (C/B)
(E) Mean Length of ASR Without a Serious Error
(F) Serious Error Distance (E/B)
(G) Word Error Rate Groups: L stands for Low; M stands for Moderate
(H) Duration

#	A	B	C	D	E	F	G	H
4	10.02	811	15	.018	49	.060	L	2:40
5	12.90	605	11	.018	48	.079	L	2:32
9	14.94	470	11	.023	36	.077	L	1:44
8	17.49	667	19	.028	31	.046	M	2:18
12	19.15	558	25	.045	20	.035	M	1:52
1	21.49	678	20	.029	31	.045	M	2:18

Table 10. Mandarin Test Clip Characteristics

#	A	B	C	D	E	F	G	H
9	18.30	233	12	.052	19	.080	L	2:13
7	19.47	223	3	.013	55	.248	L	2:02
2	23.62	254	16	.063	16	.063	L	2:19
1	26.84	260	15	.058	15	.058	M	2:23
5	27.44	289	17	.059	15	.050	M	2:26
4	29.93	284	24	.085	11	.037	M	2:28

Table 11. Arabic Test Clip Characteristics