The Impact of Evaluation Scenario Development on the Quantitative Performance of Speech Translation Systems Prescribed by the SCORE Framework

Brian A. Weiss
National Institute of Standards and Technology
100 Bureau Drive, MS 8230
Gaithersburg, Maryland 20899
301-975-4373
brian.weiss@nist.gov

Craig Schlenoff
National Institute of Standards and Technology
100 Bureau Drive, MS 8230
Gaithersburg, Maryland 20899
301-975-3456
craig.schlenoff@nist.gov

ABSTRACT
The Defense Advanced Research Projects Agency's (DARPA) Spoken Language Communication and Translation for Tactical Use (TRANSTAC) program is a focused advanced technology research and development program. The intent of this program is to demonstrate capabilities to quickly develop and implement free-form, two-way, speech-to-speech spoken language translation systems allowing speakers of different languages to communicate with each other in real-world tactical situations without the need for an interpreter. The National Institute of Standards and Technology (NIST), with support from the Mitre Corporation and Appen Pty Limited, has been funded by DARPA to evaluate the TRANSTAC technologies since 2006. The NIST-led Independent Evaluation Team (IET) has numerous responsibilities in this ongoing effort including collecting and processing training data, designing and implementing performance evaluations, and analyzing the test data. In order to design and execute fair and relevant evaluations, the NIST IET has employed the System, Component and Operationally-Relevant Evaluation (SCORE) framework. The SCORE framework is a unified set of criteria and tools built around the premise that, in order to gain an understanding of how a technology would perform in its intended environment, it must be evaluated at both the component and system levels and further tested in operationally-relevant environments while capturing both quantitative and qualitative performance data. Since an evaluation goal of the TRANSTAC program is to capture quantitative performance data of the translation technologies, the IET developed and implemented SCORE-inspired live evaluation scenarios. The two developed forms of live evaluation scenarios have unique impacts on the quantitative performance data. This paper presents the TRANSTAC program and SCORE methodology, as well as the evaluation scenarios and their influence on system performance.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Machine translation, Speech recognition and synthesis, Text analysis.*

General Terms
Design, Experimentation, Languages, Measurement, Performance.

Keywords
SCORE, TRANSTAC, Speech-to-Speech Translation System, Performance Metrics, Evaluation

1. INTRODUCTION
The Spoken Language Communication and Translation for Tactical Use (TRANSTAC) program is an advanced technology research and development program managed by the Defense Advanced Research Projects Agency[1] (DARPA) [3]. The objective of the TRANSTAC program is to demonstrate capabilities to rapidly develop and field free-form, two-way, speech-to-speech spoken language translation technologies that allow speakers of different languages to communicate with each other in real-world tactical situations without the need for an interpreter [7] [11]. To date, several prototype systems have been developed for various language domains in Iraqi Arabic, Mandarin, Farsi, Dari, Pashto, and Thai. Systems have been demonstrated on PDAs (Personal Digital Assistants), laptop-grade platforms, and compact, ruggedized laptop systems[2] with varying performance.

The primary use case of the TRANSTAC technology involves US military personnel and foreign language speakers engaging in a range of civilian and tactical dialogues. The anticipated concept of operation is that the English-speaking personnel will be trained in advance to use the technology, while it is assumed that the foreign language users will have little to no opportunity to become familiar with the system.

DARPA has funded the National Institute of Standards and Technology (NIST) to lead the evaluation of the TRANSTAC technologies, with support from the Mitre Corporation and Appen Pty Limited. As the Independent Evaluation Team (IET), NIST was tasked with capturing the required language training data, designing and implementing multiple evaluations to capture both

[1] The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

[2] Certain commercial products and software are identified in this paper in order to explain our research. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the products and software identified are necessarily the best available for the purpose.

technical performance and end-user utility assessment, and analyzing the data. This included the IET collecting technical performance data from the TRANSTAC systems under live test conditions. The IET utilized the System, Component, and Operationally-Relevant Evaluation (SCORE) framework to produce test scenarios that English and foreign language speakers used as the backbone of their dialogues between themselves while using the TRANSTAC technology [4]. These test scenarios directly impacted the metrics generated from measures captured from these test dialogues. To date, NIST has primarily evaluated English/Iraqi Arabic two-way systems along with English/Dari two-way systems.

This paper will discuss the following: Section 2 will provide background on the SCORE framework; Section 3 will present the high level concept transfer metrics that the evaluation sought to output; Section 4 will discuss the live evaluations including relevant technical performance metrics and test scenarios; Section 5 will discuss the impact of the test scenarios on the performance data[3]; Section 6 will offer a glimpse of future scenario design; and Section 7 provides conclusions.

2. SCORE METHODOLOGY

The SCORE framework is a design methodology that is built around the premise that, in order to get a true picture of how a system performs in the field, it must be evaluated at the component level, the capability level, the system level, and in operationally-relevant environments [3] [10].

SCORE is a cohesive suite of criteria and software tools employed to design performance evaluations for complex intelligent systems. It stipulates an extensive evaluation plan that is capable of both assessing technical performance through variable isolation and manipulation along with collecting end-user utility across a range of test environments.

SCORE sets itself apart from other methodologies since:

1. It can be applied to a broad range of technologies from manufacturing to defense systems
2. Elements of SCORE can be decoupled and customized based upon specific goals
3. It can evaluate a technology at varying stages of development, from conceptual to the final iteration
4. It combines the results of targeted evaluations to produce a comprehensive representation of a technology's capabilities, performance, and utility.

This framework has provided proven techniques to facilitate performance evaluations of numerous intelligent systems since it was conceived. To date, it has driven five TRANSTAC evaluations and six test events for DARPA's Advanced Soldier Sensor Information System and Technology (ASSIST) program [8] [12]. Likewise, the SCORE framework was employed to produce the initial designs for the RoboCup Rescue Virtual Robots Competition and Virtual Manufacturing Automation Competition (VMAC) [1] [2] [4].

[3] Due to DARPA restrictions, the performance data captured using these test scenarios cannot be published. Instead, this paper will focus on the approach and impact as opposed to the results.

2.1 Evaluation Goal Types

The SCORE framework has evolved over the years to define five evaluation goal types [10].

- *Component Level Testing – Technical Performance* – This evaluation type decomposes a system into components to isolate those subsystems that are critical to system operation. Ideally, all of the components taken together should include all facets of the system and yield a complete evaluation. This level of testing has occurred in past TRANSTAC evaluations where the three major components of speech-to-speech systems, Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS), were evaluated independently from one another.
- *Capability Level Testing – Technical Performance* – This type of evaluation involves identifying and isolating individual capabilities of a system and measuring their technical performance. A system can have one or more capabilities. This test type has also occurred in previous TRANSTAC evaluations when the IET designed and executed tests focused on the systems' capability of correctly translating proper names.
- *System Level Testing – Technical Performance* – This evaluation type is intended to assess the complete system, but in a controlled environment where test variables can be separated and influenced. The benefit is that tests can be performed using a combination of variables and parameters, where relationships can be determined between system behavior and these variables and parameters based upon the technical performance analysis. This test type inspired the live TRANSTAC evaluations and their driving scenarios which will be discussed in Section 4.
- *Capability Level Testing – Utility Assessments* – This evaluation type assesses the utility of an individual capability (where the complete system is made up of multiple capabilities), where utility is defined as the value the application provides to the end-user. Additionally, usability is assessed; this includes effectiveness, and user attitude towards the system. This test type also influenced several TRANSTAC evaluations where the IET captured end-user assessments of the technologies' ability to translate proper names.
- *System Level Testing – Utility Assessments* – This evaluation type assesses a system's utility and has inspired numerous live TRANSTAC technology evaluations. These include tests where Marines, Soldiers and foreign language experts provided utility feedback about the systems after using them in a range of tests.

It is important to note that even though the last two test types focus on extracting the technology users' utility assessment, it is virtually impossible to prevent the users' perceptions from being influenced by the technologies current level of technical performance. The users' utility is based upon the current state of the technology and is expected to change as the technical performance improves over future test events.

2.2 Evaluation Elements

The following evaluation elements must be identified for each goal type in order to generate relevant, reasonable, and appropriate evaluations [10].

- *Identification of the system or component to be assessed*
- *Definition of the goal/objective(s)/metrics/measures*
    - *Goal* – For a particular assessment, the goal is influenced by whether the intent of the evaluation is to inform or validate the system design.
    - *Objectives* – Evaluation objectives are used to separate evaluation concerns. These concerns also include identifying how different variables impact system performance.
    - *Metrics/Measures* – Depending upon the type of evaluation, either technical performance metrics and/or utility metrics are specified.
- *Specification of the testing environment* – Selecting the testing environment is influenced by a range of aspects including intended use-case environments, system maturity, etc.
- *Identification of Personnel* – This includes selecting the direct technology users, the test participants who will be indirectly interacting with the technology as role-players in the environment and evaluation personnel who will be directing role-players and/or capturing measures.
- *Specification of the personnel training* – All three personnel groups identified above must be given appropriate training and adequate practice time to become proficient in their test responsibilities.
- *Specification of the data collection methods* – Data capture methods, equipment, and/or instrumentation must be identified as measures and metrics are specified.
- *Specification of the use-case scenarios* – Test scenarios must be devised that are appropriate to the system or component being tested and the test end-users.

This paper is focused on the element of *Specification of the use-case scenarios* as designed and implemented within the *System Level Testing – Technical Performance* goal type. Prior to discussing these use-case scenarios, it is important to present the metrics that drive the scenario generation and implementation.

3. High Level Concept Transfer Metrics

Before discussing metrics specific to this work it is important to define both metrics and measures with respect to their usage by SCORE. Metrics are defined as the interpretation of one or more contributing elements, e.g. measures or other metrics that correspond to the degree to which a set of attribute elements affects its quality [6]. Likewise, a measure is defined as a performance indicator that can be observed, examined, detected, and/or perceived either manually or automatically [6]. For example, suppose it is desired to capture the velocity of a new vehicle under test. Examples of measures would be timing how long it takes a car to travel from one point to another and measuring the exact distance traveled. The velocity metric would be generated using the distance and time measurements where *velocity = distance/time*. Note that in some cases, a metric can be directly measurable. Using the same example, radar (or some other capture device) can directly capture the velocity of the vehicle making the measurement equal to the metric. Discussion will now follow of some of the technical performance metrics generated and/or captured during the TRANSTAC evaluations.

One of the key metrics that DARPA specified for evaluating the TRANSTAC technologies was the capture and analysis of *High Level Concept Transfer* Metrics. This suite of metrics reflects the goal of the TRANSTAC program which is the deployed use of the speech-to-speech machine translation technology to enable consistently successful communication between English-speaking and foreign language personnel [11].

Specifically, *High Level Concept Transfer* metrics consist of bilingual judges determining whether the meaning of a human-spoken utterance was conveyed during the machine translation. These metrics include the number of utterances that were successfully translated per ten minutes (with failed utterances not directly scored except for taking up time) so these metrics are assessments of both efficiency and accuracy. Additional *High Level Concept Transfer* metrics include:

- *Number of questions per 10 minutes* - Number of questions correctly translated in ten minutes as spoken by the English speaker
- *Question Percentage* - Percentage of questions that were correctly translated divided by the total number of questions asked
- *Number of attempts per question* - As spoken by the English speaker
- *Number of answers per 10 minutes* – Number of answers correctly translated in ten minutes as spoken by the foreign language speaker
- *Answer Percentage* – Percentage of answers that were correctly translated divided by the total number of answers stated
- *Number of attempts per answer* – As spoken by the foreign language speaker

It should be noted that these metrics are considered normalized since they can be computed using data from evaluation scenarios regardless of how much time it took to conduct each scenario.

Now that the evaluation type's required metrics are known, additional evaluation elements can be specified including the *Specification of the use-case scenarios*. To attain the *High Level Concept Transfer Metrics*, specific live evaluation scenarios have been designed and implemented across many of the TRANSTAC evaluations. These scenarios are discussed in the following section.

4. LIVE EVALUATIONS

A majority of each TRANSTAC evaluation features live scenarios performed by English-speaking Soldiers or Marines (also known as Subject Matter Experts or SMEs) and Foreign Language Experts (FLEs). These evaluations took place in both the lab (set up as an indoor, controlled environment where speakers remained stationary) and the field (outdoor, simulated tactical environments where the speakers were mobile and background noise was present) [7] [9] [11]. Figure 1 depicts a live field evaluation from a recent TRANSTAC test event.

Figure 1. Live evaluation in the field environment at a recent TRANSTAC test event

Both of these test environments support the capture of quantitative and qualitative data and have featured two scenario types to attain these metrics: structured scenarios and spontaneous scenarios. The following sub-sections will present both types of scenarios and how they have been employed in the TRANSTAC evaluations. A final sub-section presents how the *High Level Concept Transfer* metrics are obtained from performing the two scenario types.

4.1 Structured Scenarios

Structured scenarios were intended to prompt the SME to ask the FLE questions (or convey information, in some instances) in order to obtain information from the FLE. The concept of this scenario type is that both speakers are told exactly what pieces of information they need to collect and/or convey. However, the speakers have the latitude to phrase their question and/or statement using whatever wording they choose so they can maximize their chances of a successful dialogue. A structured scenario is composed of two separate documents: the SME version and the FLE version. The SME version contains:

- **Background** – Specific information to put the SME in the appropriate mindset. This often includes high level goals and/or a snapshot of the current state of affairs.
- **Scene** – Describes the immediate situation and specific goals.
- **Outcome** – Presents the expected result of the conversation (as stated in the structured dialogue).
- **Questions/Prompts** – Numbered list of specific pieces of information the SME is to ask of the FLE or to convey to the FLE. Note that questions with multiple numbers indicate to the SME that there are multiple concepts to be obtained from the FLE.

Likewise, the FLE version contains **Background**, **Scene**, and **Outcome** elements, but they are stated from the FLE's perspective making them unique from the SME's version. Instead of **Questions/Prompts**, the FLE version contains **Responses** comprised of informational paragraphs. These include key pieces of information in bold throughout the paragraphs. An example of a structured scenario, showing both the SME and FLE (written in English) versions, is shown in Figure 2.

The evaluation protocol for the structured scenarios begins with the SMEs and FLEs each receiving their respective versions. After reviewing their dialogues separately, the SME and FLE practice their scenario together in their native languages through an interpreter (taking the place of a TRANSTAC system). After the training session is complete, the speakers participate in the evaluation. At this point, the SME is trained on the specific TRANSTAC technology they are about to use. However, the only training the FLE receives on the technology is in the form of TRANSTAC system spoken instructions that are played by the SME immediately before the evaluation dialogue begins. As the speakers are conversing through the TRANSTAC systems according to the structured format, the SMEs are informing an IET member of the concepts they perceived from the technology. For example, if a SME asks a FLE how many children he has and the FLE responds with "I am proud to have two sons," then the SME would simply report "two sons" to the IET.

Each structured scenario was conducted by a SME/FLE pair within a ten minute window. Since the scenarios were designed with more concepts than the speakers could reasonably get through in ten minutes, the speakers never reached the end of their structured scenario dialogues.

It should be noted that the content of each structured scenario is derived from audio dialogues that were collected by the IET ahead of each evaluation [7] [9] [11]. These 20 to 25 minute interpreter-mediated dialogues occurred between Marines or Soldiers and foreign language speakers within a recording studio. These dialogues were inspired by tactically-relevant data collection scenarios that the IET developed for the data collection efforts.

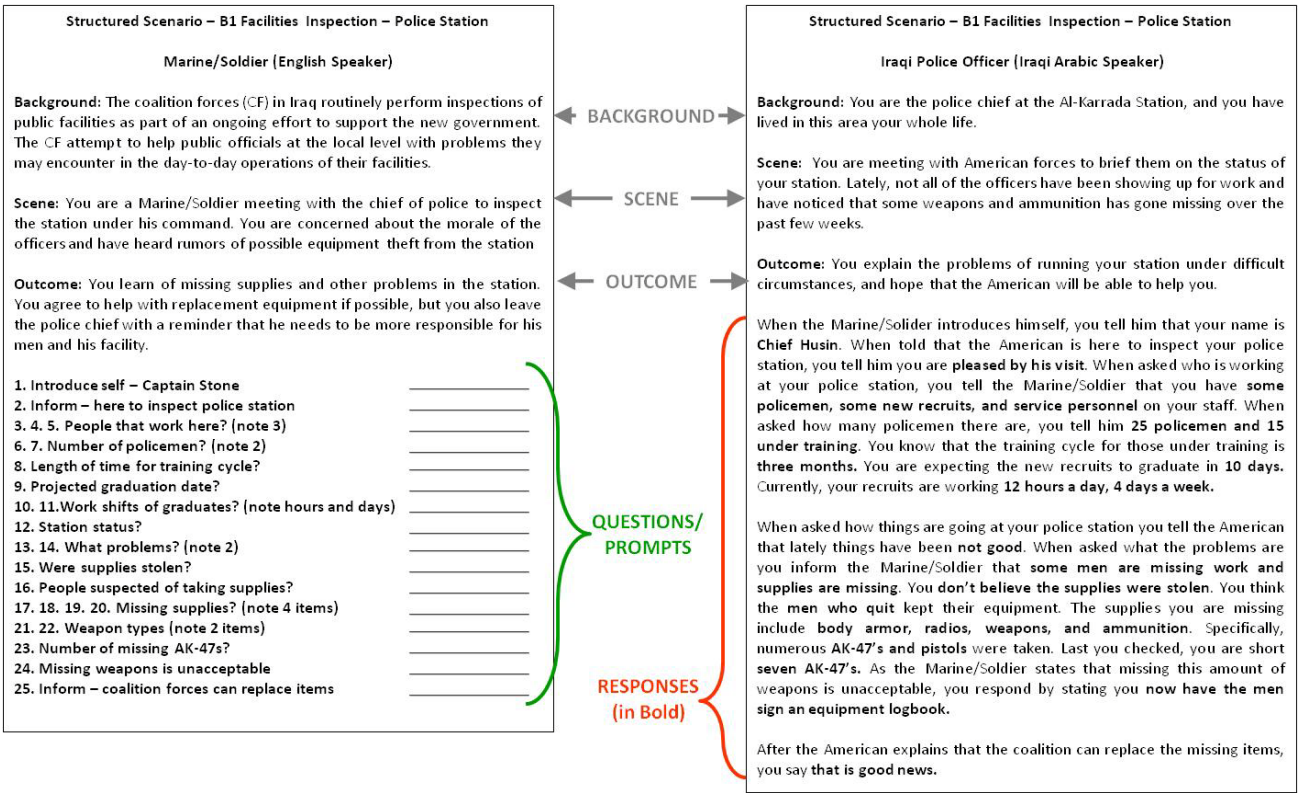

Figure 2: Structured scenario outlining a police station inspection dialogue between a Marine/Soldier and Iraqi police officer

4.2 Spontaneous Scenarios

The spontaneous scenarios provided the SMEs and FLEs with more freedom and latitude in their dialogues by not laying out specific questions and answers as compared to the structured scenarios.

A spontaneous scenario begins with specifying the overall domain (six tactical domains were commonly identified for the Iraqi Arabic and Dari systems). For each domain, multiple SME motivations were generated that included some background and situational information along with the mindset the SME should take in the conversation. Additionally, each SME motivation was paired with numerous talking points with the intent of giving the SME topics they could include in their dialogues, but not limit them to specific questions. In turn, each scenario provided the FLE with a specific motivation including some background information. These can be seen in the example shown in Figure 3.

Since these scenarios have been used in evaluations involving a single SME and multiple FLEs per conversation, it was important to generate multiple FLE motivations corresponding to a single SME motivation. Each FLE motivation was designed to be unique from one another even though they applied to the same scenario. However, the FLE motivations built upon one another where each FLE's information either supported one another, created a broader picture, or purposefully contradicted one another. An example of this can be seen in Figure 3.

An additional consideration in creating the spontaneous scenarios was the environment where they were employed. Dialogues will naturally play out differently given the environment and specific props available for the speakers to comment and discuss. Using the police station facilities inspection scenario noted in Figure 3 as an example, it is possible to have drastically different dialogues if this scenario were performed in a very old, run-down building as compared to conducting the same scenario in a brand-new, pristine facility. The more realistic the evaluation environment, the more representative the dialogues will be when driven by spontaneous scenarios.

The evaluation protocol began with each speaker being given their own motivation and unable to see their counterpart's. The SMEs and FLEs were trained separately from one another with IET assistance. Their training covered possible dialogue directions along with how to interact with one another in the simulated tactical environments set up for the evaluation. Since the scenarios required the SMEs to have a tactical background in the areas they would be discussing, the IET considered their individual experiences when devising scenario assignments. In some instances, SMEs were paired with scenarios that they were unfamiliar with so they worked with other SMEs and IET members to better understand the domains. SMEs and FLEs received comparable technology training as if they had been doing structured scenarios. The SMEs received extensive training on the systems prior to the evaluation while the FLE was played verbal instructions from the system immediately before their evaluation dialogues began.

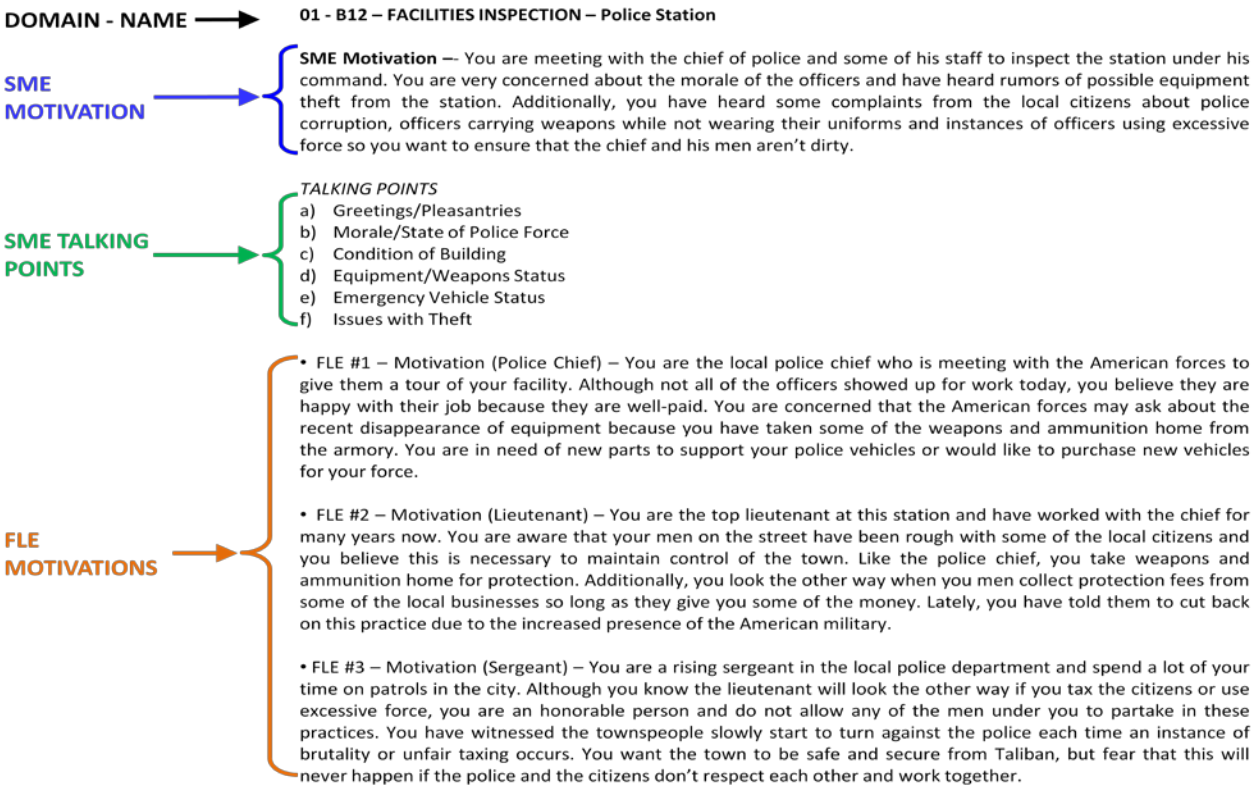

Figure 3. Spontaneous scenario outlining a police station inspection dialogue between a Marine/Soldier and Afghan police officer

The evaluations then commenced and the SMEs and FLEs role-played their dialogues. The spontaneous scenarios ran differently than the structured scenarios in that the speakers had 15 to 35 minute windows to speak based upon the evaluation schedule. Since all scenarios ran for unique amounts of time, the normalized metrics (discussed in Section 3) applied in the structured scenarios were also applicable here. This enabled the evaluation team to conduct a more "apples-to-apples" comparison of the data given the varying scenario times.

It should be noted that these scenarios stemmed directly from corresponding data collection scenarios, but were augmented to support the evaluation [9]. Like the structured scenarios, the spontaneous scenarios were also based upon the audio dialogues collected at IET-led data collections.

4.3 Metrics Generated from Scenario Data

Both the structured and spontaneous scenarios served their purpose of enabling the evaluation team to generate *High Level Concept Transfer* metrics from the live conversations between English and foreign language speakers using the TRANSTAC technologies.

Since the structured scenarios provided the IET with the concepts that were conveyed by the speakers before the evaluation began, scoring spreadsheets were devised ahead of time to support the data analysis. Once the evaluation concluded, the IET enlisted the support of ten bilingual judges to assess the accuracy of the machine translations as compared to the human speech from the SMEs and FLEs. Between three to six judges assessed each evaluation dialogue which entailed viewing and listening to the recorded scenario, noting how many attempts a speaker made to convey a concept, and scoring how successful the technologies were in translating the spoken concepts. At the conclusion of the bilingual judges' analysis, the IET averaged out all of the judgments for each scenario and calculated the metrics discussed in Section 3.

Analyzing the data from the spontaneous scenarios was similar with the exception of one time-consuming and critical difference; since the scenarios were spontaneous in nature and the concepts to be conveyed were not known ahead of time, the IET had to transcribe the evaluation conversations and identify the concepts that the speakers were attempting to transfer. Ultimately, both scenarios produced the same desired metrics to assess this aspect of the TRANSTAC technologies' technical performance.

5. SCENARIO IMPACT ON METRICS

Both the structured and spontaneous scenario types impacted the evaluation dialogues which in turn, impacted the *High Level Concept Transfer* metrics. The following sub-sections present the specific impacts and how these affected the metrics.

5.1 Impacts

When conducted across multiple evaluations, the structured scenarios allowed the following with each having unique effects.

- The same structured scenarios using the same concepts were used across multiple evaluations.
  EFFECT – Direct technical performance comparisons were drawn across multiple technologies over multiple evaluations enabling more "apples-to-apples" assessments.
- SMEs and FLEs did not need firsthand knowledge of a particular scenario to be effective as long as they had sufficient training to become familiar with the concepts.

EFFECT – It was easier to obtain some repeatability across multiple speakers who performed the same scenarios.

- SMEs and FLEs were forced to attempt specific concepts, some of which were not easily understood by the technology.
  EFFECT – Technologies had to attempt varying targeted and challenging vocabulary that would have not been otherwise attempted.
- SMEs and FLEs were given little flexibility in their dialogues so it was easy for them to become disengaged in their conversations.
  EFFECT - Speakers were more prone to speaking less-naturally leading to a decrease in the ability of the technology to recognize their speech.
  EFFECT – Speakers were prone to reading concepts verbatim from the scenarios, as opposed to rephrasing, even when they had to repeat them due to miscommunications.

The spontaneous scenarios counteracted some of the negative consequences of the structured scenarios while producing some other effects, as well. These scenarios allowed the following producing the noted affects.

- Speakers used the system in the anticipated manner in which it would be deployed in more relevant, use-case environments.
  EFFECT – The output metrics provided a more representative gauge of how the system would perform in actual use-case environments.
- The same scenarios using the same talking points could be used across multiple evaluations but would still ultimately produce very distinct dialogues.
  EFFECT - It would be very challenging to make direct technical performance comparisons across multiple evaluations.
- SMEs and FLEs had great flexibility in their dialogues as long as their responses stay consistent and they remain within the scope of the scenario.
  EFFECT – The speakers made the scenarios "their own" thereby becoming more engaged and enthusiastic.
- SMEs must have firsthand knowledge of a scenario's tactical domain to effectively role-play the conversation during the evaluation.
  EFFECT - All of the dialogues were unique since they were based upon the SMEs' distinct experiences.
- SMEs and FLEs must improvise during their conversations in the event that the TRANSTAC systems were having difficulties with specific areas of dialogue.
  EFFECT – Dialogues easily stalled if the speakers did not change their wording or conversation direction based upon the systems' vocabulary capabilities.

5.2 Impact Analysis

After analyzing the *High Level Concept Transfer* metrics from multiple evaluations that were supported by structured and spontaneous scenarios, the following observations were made:

- On average, scores across all of the *High Level Concept Transfer* metrics were lower for those scenarios that forced the speakers to use specialized vocabulary, i.e. the scenarios performed within the medical domain scored lower as compared to the overall averages
- On average, the *Number of Attempts per Question* and *Number of Attempts per Answer* were higher for evaluations supported by the spontaneous scenarios as compared to the structured scenarios
- On average, the *Number of Questions per 10 minutes* and the *Number of Answers per 10 minutes* were lower for evaluations supported by the spontaneous scenarios as compared to the structured scenarios
- On average, the *Question Percentage* (number of questions correctly translated over the number of total questions asked) and the *Answer Percentage* were lower for those evaluations supported by the spontaneous scenarios as compared to the structured scenarios
- Since the FLEs had more flexibility in the spontaneous scenarios and weren't constrained to specifying multiple concepts per response, as they were in the structured scenarios, the average ratio of questions to answers was lower in this scenario type as compared to the structured scenarios resulting in less answer opportunities

It is important to note that the scenarios were not the only significant factor contributing to the disparity in results of metrics when applied to data from structured and spontaneous scenarios. All of the *High Level Concept Transfer* metrics captured using the structured scenarios have resulted from evaluations testing the Iraqi Arabic (IA) TRANSTAC systems. In contrast, the spontaneous scenarios have only been applied to the most recent evaluation which tested the Dari versions of the TRANSTAC technology. Additionally, the technology developers have had access to the IA data for a much longer period of time as compared to the Dari data. Also there is much more IA conversation data available to support training and development efforts as compared to the limited amount of available Dari data.

6. FUTURE EFFORTS

The IET is expecting to deploy another round of spontaneous scenarios to support the October 2009 evaluation. The overriding factor in the selection of spontaneous scenarios over structured scenarios is that the spontaneous scenarios enable the speakers to use the system in the expected manner in which it would ultimately be deployed, thereby providing an indication of the technology's current performance level under these conditions. This is critical considering it is desired to provide this technology to Soldiers and Marines operating within tactical environments in the near future.

The October 2009 evaluation will test the TRANSTAC research teams' two-way, English/Pashto systems to capture both technical performance (including the discussed *High Level Concept Transfer* metrics) and end-user qualitative assessments. The IET is exploring ways to augment the spontaneous scenario including the addition of suggested pieces of information to capture, i.e. presenting the SME with structured scenario-like prompts that they could optionally ask. Ultimately, the SME would still be free to take the conversation in any direction within the scenario's scope, but would have the fallback option to ask some (or all, at their discretion) of the IET-specified questions. However, the FLEs' scenarios would remain unchanged. Their dialogue would still be governed by their scenario-driven motivation where they would respond with answers relevant and consistent with the scenario.

7. CONCLUSION

SCORE has proven to be an invaluable evaluation design generation tool in formulating appropriate performance tests for DARPA's TRANSTAC technologies. This framework inspired the creation and implementation of the structured and spontaneous scenarios across multiple test events. Each scenario type has yielded vast amounts of data to support the suite of *High Level Concept Transfer* metrics necessary to the IET's evaluation. To date, SCORE has driven the development of 11 DARPA evaluations including six for the ASSIST program and five for the TRANSTAC program along with providing design inspiration to the VMAC and RoboCup Rescue Virtual Robot competitions. Based upon the success of these evaluations including the comprehensive levels of data generated, the IET envisions using this framework to support future evaluations of advanced technologies and other intelligent systems under test.

8. ACKNOWLEDGMENTS

The authors would like to acknowledge the DARPA TRANSTAC program manager, Dr. Mari Maeda, and the members of the NIST IET for their continued support.

9. REFERENCES

[1] Balakirsky, S., Carpin, S., Dimitoglou, G., and Balaguer, B. 2009, "From Simulation to Real Robots with Predictable Results: Methods and Examples," in *Performance Evaluation and Benchmarking of Intelligent Systems*, New York: Springer Science & Business Media, 2009, pp. 113-137.

[2] Balakirsky, S. and Madhavan, R. 2009. Advancing Manufacturing Research Through Competitions. In Proceedings of the SPIE Defense Security and Sensing Conference (Orlando, Florida, USA, April 13 – 17, 2009).

[3] DARPA. 2009. Spoken Language Communication and Translation System for Tactical Use (TRANSTAC). http://www.darpa.mil/IPTO/programs/transtac/transtac.asp

[4] Intelligent Systems Division – National Institute of Standards and Technology. 2009. Measurement Science for Intelligent Manufacturing Robotics and Automation Program. http://www.nist.gov/mel/isd/si/msimra.cfm

[5] Intelligent Systems Division – National Institute of Standards and Technology. 2009. System, Component and Operationally-Relevant Evaluations (SCORE). http://www.isd.mel.nist.gov/projects/score/

[6] Schlenoff, C., Steves, M.P., Weiss, B.A., Shneier, M. and Virts, A. 2007. Applying SCORE to Field-Based Performance Evaluations of Soldier Worn Sensor Technologies. Journal of Field Robotics – Special Issue on Quantitative Performance Evaluation of Robotic and Intelligent Systems, vol. 24 (Sept 2007), pp. 671 – 698.

[7] Schlenoff, C., Weiss, B.A., Steves, M.P., Sanders, G., Proctor, F., and Virts, A. 2009. Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies. To Appear In Proceedings of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Workshop (Gaithersburg, Maryland, USA, September 21 – 23, 2009).

[8] Schlenoff, C., Weiss, B.A., Steves, M., Virts, A, and Shneier, M. 2006. Overview of the First Advanced Technology Evaluations for ASSIST. In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop (Gaithersburg, Maryland, USA, August 21 – 23, 2006).

[9] Weiss, B.A. and Menzel, M. 2009. Development of Domain-Specific Scenarios for Training and Evaluation of Two-Way, Free Form, Spoken Language Translation Devices. In Proceedings of the 2009 International Test and Evaluation Association (ITEA) Symposium (Baltimore, Maryland, USA, September 28 - October 1, 2009).

[10] Weiss, B.A. and Schlenoff, C. 2008. Evolution of the SCORE Framework to Enhance Field-Based Performance Evaluations of Emerging Technologies. In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop (Gaithersburg, Maryland, USA, August 19 - 21, 2008).

[11] Weiss, B.A., Schlenoff, C., Sanders, G.A., Steves, M.P., Condon, S., Phillips, J., and Parvaz, D. 2008. Performance Evaluation of Speech Translation Systems. In Proceedings of the 6th edition of the Language Resources and Evaluation Conference (Marrakech, Morocco, May 28 – 30, 2008).

[12] Weiss, B.A., Schlenoff, C., Shneier, M., and Virts, A. 2006. Technology Evaluations and Performance Metrics for Soldier-Worn Sensors for ASSIST. In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop (Gaithersburg, Maryland, USA, August 21 – 23, 2006).