Michael, E. B., Massaro, D., & Perlman, M. (in press). What's the bottom line? Development of and potential uses for the Summary Translation Evaluation Tool (STET). *The Next Wave.*

What's the bottom line?

Development of and potential uses for the Summary Translation Evaluation Tool (STET)

Erica B. Michael, PhD, Dominic W. Massaro, PhD, Marcus Perlman, MS

Language analysts in the intelligence community (IC) confront huge amounts of foreign language material, some highly relevant to national security requirements and some of lesser value. Language analysts cannot provide full translations of everything they process, nor do other analysts have the time or need to read full translations. Summary translation enables language analysts to identify, distill, and present English translations of the important information contained in the original foreign language material, thereby efficiently communicating the most crucial information.

The term summary translation[1] can be used to cover a broad range of tasks that vary in their purposes and skill demands, from documenting essential elements of information contained in a single source item to writing a personality profile or situation assessment synthesizing information from many source items (Michael, Bailey, Gannon-Kurowski, & Pinckney, 2007). Despite this variability, a common attribute of summary translations produced within the IC context is that they are typically *targeted* summaries written in response to intelligence needs, i.e., customer-specified requirements or requests for information. In other words, writers may be looking for information about specific topics or answers to specific questions, rather than attempting to summarize all of the main points of the source item as they would in a *generic* summary.

The Summary Translation Evaluation Tool (STET) was created primarily for assessment of targeted summaries, which are uncommon in commercial and academic environments and therefore rarely studied and ill-understood. The STET was designed not only to help researchers develop a deeper understanding of targeted summary translation, but also to establish a standard for summary translations and to provide language analysts with a vital tool for both assessing and improving summary translation performance via standardized quality control (QC) and enhanced training.

WHAT IS THE STET?

The STET is a computerized form that offers a standardized framework for evaluating summary translation products.[2] The heart of the STET is a set of rating scales to assess summaries along six different dimensions: Significance, Completeness, Accuracy, Omission of Irrelevant Information, Organization, and Writing. Each dimension is described in Figure 1. These dimensions were designed to cover all of the important elements of a summary's content, structure, and style, but the relative importance of each dimension may depend on the purpose for which the summary is written.

The STET also includes a description of the source item(s) on which the summary is based. Although users of the STET are instructed not to adjust their ratings based on the difficulty of the material, the Source Item Description identifies features of the material that may be especially challenging and helps provide context for the STET ratings.

As described in the STET user's manual, the Source Item Description and Summary Translation Assessment are "analogous to the difficulty and execution in an Olympic dive; one must describe both the difficulty of the task and the skill with which it is executed in order to make a meaningful evaluation" (p. 3).

INSERT COPY OF THE STET ABOUT HERE (SEE ATTACHMENT)

---

[1] The term *gisting* is sometimes used synonymously with *summary translation*, but we will not use *gisting* here because it often refers to a process that would not typically be evaluated with the Summary Translation Evaluation Tool (STET). For example, in some operational environments analysts use *gisting* to refer to the process of making brief notes about the content of an item for triage purposes, and *summary translation* to refer to a more formal summarization process.

[2] The STET has also been adapted for evaluation of other translation products. For example, one operational organization has created a spin-off called the Language Product Evaluation Tool (LPET), which can be used to evaluate summary translations, verbatim translations, or hybrids (in which some material is summarized and some is translated verbatim).

Figure 1 shows the STET form along with descriptions of each major component. In the interactive version of the tool, pop-up windows provide more detailed information than is available on the one-page static form. For the Summary Assessment section on the right-hand side of the form, each pop-up window describes the fundamental question that is addressed in the rating scale, often indicating what qualities a summary should possess to receive a high rating on that scale. Each pop-up window also provides labels for the end-points of the rating scale. The labels are tailored to each dimension; for example, the Organization scale ranges from "The summary is extremely poorly organized" to "The summary is extremely well organized." Finally, because the STET is intended to be consistent with the eight analytic standards issued by the Director of National Intelligence (McConnell, 2007; see Table 1), each pop-up window lists the particular standards that are addressed by that dimension.

Table 1. ODNI standards of analytic tradecraft (McConnell, 2007)

1. Properly describes quality and reliability of underlying sources
2. Properly caveats and expresses uncertainties or confidence in analytic judgments
3. Properly distinguishes between underlying intelligence and analysts' assumptions and judgments
4. Incorporates alternative analysis where appropriate
5. Demonstrates relevance to US national security
6. Uses logical argumentation
7. Exhibits consistency of analysis over time, or highlights changes and explains rationale
8. Makes accurate judgments and assessments

POTENTIAL USES OF THE STET

Quality control

Language analysts and supervisors throughout the IC recognize the importance of QC, but there is currently little standardization of QC procedures. In addition to its critical function of ensuring that products are accurate, QC serves as a mechanism for training and providing feedback to language analysts. QC may also contribute to record-keeping and help supervisors determine work assignments. The STET is designed to facilitate all of these aspects of QC.

Help for the QC provider – a more efficient and effective means of conducting QC

One of the most critical functions of QC is ensuring that intelligence products are of high quality. QC is important not just for a final report, but also for the translations, summaries, and other possible steps that may be completed along the way; if the original source material is translated and/or summarized inaccurately, there is a high risk that the final report will contain incorrect information.

QC is seen as especially crucial for junior analysts who may have limited experience with the target and/or language. However, most professionals realize that even highly seasoned experts can benefit from having their work reviewed by colleagues. During structured interviews conducted by our summary translation research team, nearly all language analysts noted that "nothing goes out the door without being seen by at least two pairs of eyes."

Because of the vast amount of material that is processed every day, QC places a heavy burden on the most experienced analysts. One goal of the STET is to facilitate the process of conducting QC. Although an initial time investment may be required for QCers to learn about the STET and become accustomed to using it, the STET can ultimately make QC faster and more effective by providing a standardized framework for evaluating summary translations.

Help for the QC recipient – more detailed and useful feedback

The "checking" aspect of QC emphasizes forward movement in the sense that each piece of work is checked and passed forward to the next person in the chain. For example, a language QCer may check a translation and pass the corrected version forward to a reporter.

Ideally, QC also involves a "backward" step in which feedback is provided to the language analyst who wrote the initial translation or summary. Such feedback is vital for improving the junior analysts' language skills and helping them to avoid similar mistakes in the future. Unfortunately, the fast operational tempo sometimes makes it difficult for QCers to provide feedback in a timely manner.

The structure of the STET will allow QCers to generate feedback at the same time that they are checking the work, thus helping both the QC provider and recipient, and the multidimensional nature of the STET will ensure that the feedback is detailed enough to help language analysts pinpoint specific areas for improvement.

Help for the supervisor – a mechanism for tracking progress

Over time, the STET will help supervisors and managers keep track of strengths and weaknesses in order to determine work assignments and note opportunities for targeted training for individuals or groups. For example, a supervisor who is using the STET to track an analyst's progress may note that the analyst performs very well with Level 2 material but less well with Level 3 material; this type of pattern can be useful in making appropriate work assignments. In addition, the detailed nature of the Source Item Description may reveal that the analyst excels in the face of certain challenges but struggles with others, allowing for identification of individually tailored professional development activities. Aggregated STET data may also help managers to determine whether an entire shop's performance is affected by factors such as new software tools, new mentoring programs, or changes to the physical workspace (see "Workspace Tips" sidebar, page X).

Training

Not only will on-the-job training benefit from the STET via improved QC and feedback, but classroom training will also be able to capitalize on the STET. Perhaps most critically, the STET will provide instructors with a coherent framework for teaching students about the components of a good summary. Instructors can also use the STET to provide standardized feedback on student assignments, which will make the grading process more efficient for instructors and more useful for students. Using the STET in the classroom will also help students become accustomed to the way they will be evaluated on the job.

The Source Item Description section of the STET may also be helpful for instructors in guiding the selection of texts that are at an appropriate level for the class and that present students with particular challenges that are relevant to the lesson.

Research

One of the goals of CASL's research on translation is to understand the cognitive and procedural processes involved in summary translation, and to apply that understanding to improve the performance and training of language analysts. With respect to both aspects of this aim—understanding and application—evaluation is an essential component. The STET will provide researchers with a valuable mechanism for evaluating the summary translations that are produced in experiments. The STET is a powerful tool for experimentation because it allows the researchers to examine summarization performance along a variety of dimensions.

In one CASL experiment using the STET, we are examining the ways in which varying amounts of time pressure impact the quality of summary translations and the strategies that language analysts use to create them. In this experiment, we ask language analysts to summarize a different foreign language text in each of three time conditions: 2 hours, 1 hour, and ½ hour. (Order of the time conditions and assignment of text to condition are counterbalanced across participants.)

With a holistic rubric, we would only be able to determine whether summary translations were "better" in one condition than in another. With the STET, however, we can look at the effects of time pressure on different aspects of the summarization process. For example, we might see that Significance is relatively unaffected by time pressure if language analysts prioritize the need to identify the critical intelligence value of the source item; Completeness, on the other hand, might suffer under extreme time pressure if the language analyst does not have sufficient time to include all important details. Similarly, Organization might be relatively stable but Writing might be vulnerable to time pressure when language analysts do not have time to proofread or check their work.

This detailed level of analysis enabled by the STET will help researchers better understand the various components of the summary translation task and guide the development of interventions to help analysts maintain key components of summary quality under trying conditions.

DEVELOPMENT OF THE STET

As part of CASL's first experimental study of summary translation, our research team developed a "holistic" rubric, which assigned a single qualitative rating to each summary. The scale was developed using a modified empirically-based binary-boundary approach (e.g., Upshur & Turner, 1995), meaning that we relied on collaboration and consensus of qualified professionals using an iterative process of categorizing and characterizing salient features to arrive at descriptors for each level of proficiency (Turner, 2007). This type of scale is probably similar to informal evaluations used in the IC

and is typically fairly quick to use. However, a holistic evaluation provides only a single rating of the summary, and two summaries could receive the same rating for very different reasons (e.g., one due to poor comprehension of the source item and one due to poor English writing skills).

We ultimately decided that an "analytic" rubric would be more useful for both experimental and applied purposes, as described above, i.e., summary translation evaluation via an analytic rubric returns much more informative feedback about performance, allowing for more powerful experimentation as well as more individualized on-the-job evaluation and training. The current version of the STET was developed as a collaborative effort between CASL researchers and our USG colleagues, capitalizing on scholarly literature, scientific methods, and the operational expertise of many language analysts.

Characteristics of an analytic rubric

Our development of an analytic rubric had the goal of making transparent the component processes involved in summary translation while also ensuring that the STET would be easily understood and used. Preliminary effort sought to derive a set of unidimensional evaluatory elements, and was informed both by the existing scientific literature and by an analysis of the original holistic rubric. Following standards in educational and psychological measurement (e.g., Crocker & Algina, 2006), we strove to produce an analytic rubric characterized by the following qualities:

- All elements worthy of evaluation are included.
- Each element is unidimensional in that it cannot be further separated or partitioned. (Given this quality, we refer to each element as a dimension.)
- Ratings are distinct, comprehensive, and descriptive in that they cover the range of expected performance.
- Each element of the rubric communicates clearly to the user.
- The rating score on each element covers the range of performance, perhaps in the range of 3-7 levels.

Characteristics of a good summary

To develop the appropriate dimensions for the STET, we needed to identify the most important components of summary translation based on existing science and operational needs. From a scientific perspective, we began by examining literature in fields such as language processing, memory, translation, reading comprehension, and spoken-discourse comprehension. To determine what constitutes a good summary for operational purposes, we conducted structured interviews with language analysts, intelligence analysts, QCers, and instructors. We asked interviewees to describe the ways in which they write or use summary translations on the job and what they look for in a good summary translation. In addition to these individual structured interviews, focus groups were convened to determine the qualities that users would look for in an evaluation tool and to solicit comprehensive feedback on preliminary drafts of the STET.

One of the most important characteristics of a good summary is that it accurately reflects the meaning of the source text. According to Jonassen, Beissner, and Yacci (1993), the meaning of a source text comes from its structure – the way its propositions are related to each other and organized. In other words, the reader must derive structural knowledge of the text and be able to integrate each successive part of the text with his or her prior representation of it. Related to this point, learning and understanding of the text take place by assimilating new knowledge with prior background knowledge.

For a language analyst to produce a quality summary, he or she must have (1) sufficient language proficiency to achieve a discourse-level understanding of the text and not just a word-by-word glossing; and (2) adequate background knowledge to assimilate the text with prior structural knowledge. Preliminary CASL research on summary translation demonstrated that an insufficient discourse-level understanding of the text often led to gross misunderstandings and eventually to inadequate summaries (Michael et al., 2007). A solid discourse-level understanding of the text is assessed most directly by the Accuracy dimension of the STET. In addition, a thorough understanding of the text and its relation to the relevant background knowledge are also necessary for the reader to determine and explain how and why the text relates to intelligence needs, as assessed by the Significance dimension of the STET.

Endres-Niggemeyer (1998) emphasizes the importance of representing and understanding discourse via schemata (structured groups of concepts used to organize knowledge). This implies that summarizers must be able to identify the schematic elements of the text and the relationships and actions between them. Some of these elements will be crucial to the summary, and the success of their identification should be evaluated by the rubric. For example, does the summarizer correctly identify the relevant participants, their roles in the text, and the actions taking place? Identification of the key pieces of information in the text is assessed by the Completeness dimension of the STET.

Another fundamental skill required for summarization is information reduction. Ultimately, in a quality summary, only the relevant information should remain, and the irrelevant information should be discarded, as measured by the Omission of Irrelevant Information dimension of the STET.

Finally, the summary translation must be appropriately communicated, which might be evaluated as having macrostructural and microstructural dimensions (Organization and Writing in the STET). The former is clearly reflected in the organization of the summary and the latter in its grammar. Important macrostructure components include clear, coherent

overall organization and a structure that makes evident how the summary responds to the relevant information need(s). Important microstructure components include proper grammar, spelling, and word choice, as well as a writing style that is clear at both the sentence and clause level.

Properties of the rating system

Once the dimensions were established, we had to decide on other properties of the rating system, including the appropriate number of points on the rating scale and the descriptors of the different levels.

We ultimately decided to use a 5-point rating scale. Although some of the potential users we consulted felt that a smaller number of points would make the scale faster and easier to use, research has demonstrated that reliability and validity are better in 5- to 7-point scales than in scales with fewer points (e.g., Bandalos & Enders, 1996; Dawes, 2008; Jenkins & Taber, 1977; Lissitz & Green, 1975). Also, respondents tend to avoid using the endpoints of a scale (Beal & Dawson, 2007), so having a 4-point scale could potentially concentrate most of the responses on only two points, which may not be sensitive enough to detect subtle differences in summary quality.

Another important decision was how to label the five points of the scale. We initially attempted to write a detailed description of each rating of each dimension of the STET. We discovered, however, that most of the important information appeared in the descriptor for the highest rating, which listed the characteristics required for a high quality summary. By putting all of the desired characteristics up front after each question, the endpoints of the scale emerged naturally without the need for descriptions of the intermediate ratings. This decision was validated by research suggesting that the labels for intermediate scale intervals are not as critical as the choice of endpoint labels (e.g., Gannon & Ostrom, 1996; Klocktars & Yamishi, 1988).

Refining the STET

Once the dimensions and rating system were established, we used several approaches (some of which are still in progress) to refine the STET.

Rigorous practical testing

Since a major aim of rubric development is the application of the rubric to the training and evaluation of language analysts working within government agencies, practical testing by representatives of those agencies is crucial. This "beta testing" will allow agencies to report experiences using the rubric and to provide feedback so we can make the rubric maximally user-friendly and useful for their needs. For example, it will likely be valuable to build into the rubric a degree of modularity, so that particular dimensions can be added or omitted as needs dictate. Similarly, the size of the rating scale might be collapsed or expanded according to practical needs. In coordination with practical testing, statistical testing will be used to confirm that adapted versions of the rubric are valid, sensitive, and reliable.

Rigorous statistical testing

Rigorous statistical testing is required to examine the validity, sensitivity, and reliability of the STET against gold standards, i.e., summaries pre-established by experts to represent certain levels of performance along the different dimensions. This testing will allow us to determine if variations in each aspect of summary quality are appropriately reflected in ratings for the corresponding dimension (validity) and if the STET adequately assesses the full range of performance along each dimension (sensitivity). Statistical testing will also help to determine if the dimensions are treated independently or if, for example, grammatical errors affect Accuracy scores as well as Writing scores. Lastly, the STET will be tested thoroughly to establish its consistency across users and conditions (reliability). CASL researchers are currently conducting a set of experiments to accomplish these goals.

CONCLUSIONS

The STET is the result of a needs-based approach to research in which a multidisciplinary team of scientists collaborated with members of the operational workforce to develop a product that addresses both operational and scientific problems. This analytic rubric for evaluating summary translations will benefit language analysts, QCers, and their managers, resulting in better reports and better language analysts. The STET will also allow CASL scientists to deepen our understanding of the summary translation process so we can continue to conduct rigorous research to enhance language performance in the IC.

REFERENCES

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education, 9,* 151-160.

Beal, D. J., & Dawson, J. F. (2007). On the use of Likert-type scales in multilevel data: Influence on aggregate variables. *Original Research Methods, 10,* 657-672.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* Mason, OH: Thomson Wadsworth.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 50,* 61-77.

Endres-Niggemeyer, B. (1998). *Summarizing information.* Berlin: Springer-Verlag.

Gannon, K. M., & Ostrom, T. M. (1996). How meaning is given to rating scales: The effects of response language on category activation. *Journal of Experimental Social Psychology, 32,* 337-360.

Jenkins, G. D., Jr., & Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62,* 392-398.

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Jonassen, D. H., Howland, J., Moore, J., & Marra, R. M. (2003). *Learning to solve problems with technology: A constructivist perspective* (2nd ed.). Columbus, OH: Merrill/Prentice-Hall.

Klocktars, A. J., & Yamishi, M. (1988). The influence of labels and position in rating scales. *Journal of Educational Measurement, 25,* 85-96.

Lab for Language Sciences. (2008). *The Summary Translation Evaluation Tool: What it is and how to use it.* [User's manual]. College Park, MD.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60,* 10-13.

McConnell, J. (2007). *Intelligence community directive number 203: Analytic standards.*

Michael, E. B., Allison, T., Danks, J., de Terra, D., Massaro, D., Donavos, D., Graham, K., & Klavans, J. (2007, May). *Does verbatim translation help summary translation?* Paper presented at the 6th International Symposium on Bilingualism, Hamburg, Germany.

Michael, E. B., Bailey, B., Gannon-Kurowski, S., & Pinckney, K. (2007). *Technical report M.19: Description of summary translation task requirements for specific jobs and the uses of the summaries.* College Park, MD: University of Maryland Center for Advanced Study of Language.

Turner, J. (2007). Report on working session: Rubric development and rater orientation. In B. Bailey, J. Turner, K. Pinckney, D. de Terra, & E. Michael, *M.16: Compilation of milestones examining evaluation of translation summaries.* College Park, MD: University of Maryland Center for Advanced Study of Language.

Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal, 49,* 3-12.

Summary Translation Evaluation Tool

Item ☐ Language Analyst ☐ QCer 1 ☐ QCer 2 ☐

A. Source Item Description ◯ voice ◯ graphic ◯ both

A.1. Language Level (Select overall level of source; mark all characteristics of significance in source.) ☐ 1 2 3 4

- ☐ cultural information
- ☐ diagrams, charts, graphs
- ☐ greater than average length
- ☐ high density of information
- ☐ highly technical subject matter
- ☐ inference based on overt info
- ☐ intentional deception
- ☐ lack of continuity
- ☐ meaning beyond the literal
- ☐ multiple objects or concepts
- ☐ rhetorical devices
- ☐ shared knowledge
- ☐ spatial relationships
- ☐ telling out of sequence

A.2. Complicating Mode Factors (Mark all of significance in source.)

- ☐ communicants speaking over one another
- ☐ corrupt source
- ☐ dialect
- ☐ distortion
- ☐ elliptical or telegraphic style
- ☐ heavy accent
- ☐ more than one language or dialect or alphabet
- ☐ non-standard abbreviations or specialized terminology
- ☐ non-standard colloquialisms or slang
- ☐ omissions
- ☐ one-sided conversation
- ☐ poor grammar
- ☐ poor handwriting
- ☐ poor spelling
- ☐ rapid speech
- ☐ sudden changes in subject
- ☐ typographical errors
- ☐ urgency (need for time-sensitive processing)

A.3. Impact of Complicating Mode Factors (Mark one.)
◯ none ◯ inconsequential ◯ moderate ◯ considerable ◯ extensive

B.8. QCer 2 Comments (Use appropriate handling and classification markings, if needed.)

B. Summary Assessment (Mark a number for each factor; add comments.)

B.1. Significance: How well does the summary relate the "so what" of the source item to requirements? H 1 2 3 4 5 NA

B.2. Completeness: How much of the essential information is covered in the summary: H 1 2 3 4 5
who? what? when?
where? why? how?
relevant background? analytic comment?

B.3. Accuracy: How much of the information in the Summary is accurate? H 1 2 3 4 5

B.4. Omission of Irrelevant Information: How well does the summary omit irrelevant information? H 1 2 3 4 5

B.5. Organization: How well organized is the summary: H 1 2 3 4 5
"bottom line" up front? logical organization?
well-structured paragraphs?

B.6. Writing: How well does the summary follow conventions for: H 1 2 3 4 5
grammar? spelling?
punctuation? word usage?
date, time, transliteration, etc.?

B.7. QCer 1 Comments on Summary Translation (Use appropriate handling and classification markings, if needed.)

Help | Reset Part A | Reset All | Reset Part B

Language Level—Users determine the ILR level of the source item and indicate which characteristic(s) of the item contributed to that level. In the electronic version of STET, the language levels and characteristics include pop-up windows with descriptions and examples.

Complicating Mode Factors—The presence of these factors can make an item more difficult to understand, translate, or summarize. Some factors are specific to one modality; other factors, such as colloquialisms or poor grammar, can be found in both voice and graphic items.

Impact of Complicating Mode Factors—Users assess the degree to which the complicating factors impact the ability to understand, translate, or summarize the item. For example, typographical errors may be relatively inconsequential in one text but may render another virtually incomprehensible.

Written Comments—Users are encouraged to provide written comments. Ratings on a single dimension (compared to an overall rating) can reflect specific problems that might benefit from different types of interventions. For example, a Writing score could be poor because the summary is full of typographical errors or because the writer is a non-native speaker of English who does not have an adequate grasp of English grammar.

Significance—The summary should clearly demonstrate why the source item is relevant to national security and should indicate how the relevant information relates to what is already known about a particular requirement. Users may indicate that the dimension is not applicable if, for example, a generic summary is required rather than a targeted summary.

Completeness—Users address the degree to which the summary contains all of the relevant information. In addition to presenting the facts that are explicitly stated within the source item, a good summary may need to include explanatory facts available elsewhere and analytic comments.

Accuracy—Many consider accuracy to be the most fundamental component of STET. If a summary receives a low rating for accuracy, the summary will be of very limited value even if it receives high ratings for all of the other dimensions.

Omission of Irrelevant Information—A critical feature of summarizing is efficiently communicating the most important information in the source material. Because of the targeted nature of summary translation within the IC, the source material may often contain a great deal of information that is not relevant to national security.

Writing—Poor grammar, spelling, and punctuation can obscure the message in an otherwise good summary. This dimension takes into account conventions for reporting dates, times, and transliterations of foreign names—particularly important because foreign names can be spelled many different ways in English.

Organization—A good summary must be organized in such a way that the message is communicated clearly. One of the most highly valued organizing principles within the IC is "bottom line up front." Sometimes language analysts will need to alter the original organization of the source material to convey the information in a way that most directly addresses the information need.