How to Write Items for a Construct-Based Proficiency Test for MT Evaluation

Martha Herzog

This document provides a practice exercise for writing multiple-choice test items to evaluate machine translation quality.

Contents

PART ONE: WHAT THE PROFICIENCY SCALES SAY ABOUT DIFFICULTY LEVELS OF TEXTS

We will use excerpts from Jim Child's definitions of text modes, Ray Clifford's explanation of author purpose, the ILR scale descriptor, and the ACTFL scale descriptor.

Level 1

Orientational mode.

Author purpose: Orient by communicating main ideas.

ILR descriptor: Very simple connected written material. Simple language containing only the highest frequency structural patterns and vocabulary. Texts may include simple narratives of routine behavior, highly predictable descriptions of people, persons, or things; and explanations of geography and government such as those simplified for tourists.

ACTFL version: Simple, predictable, loosely connected texts. Readers rely heavily on contextual clues. . . can most easily understand information if the format of the text is familiar, such as in a weather report or a social announcement. . . understand texts that convey basic information such as that found in announcements, notices, and online bulletin boards and forums. These texts are not complex and have a predictable pattern of presentation. The discourse is minimally connected and primarily organized in individual sentences and strings of sentences containing predominantly high-frequency vocabulary.

Level 2

Instructional mode.

Author purpose: Instruct by communicating structured factual information.

ILR descriptor: Straightforward familiar factual material. Uncomplicated, but authentic prose on familiar subjects that are normally presented in a predictable sequence which aids the reader in understanding. Texts may include descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, social notices, formulaic business letters, and simple technical material written for the general reader. Predominantly in straightforward, high frequency sentence patterns.

ACTFL version: Conventional narrative and descriptive texts, such as expanded descriptions of persons, places, and things and narrations about past, present, and future events. These texts reflect the standard linguistic conventions of the written form of the language in such a way that readers can predict what they are going to read. Readers understand the main ideas, facts, and many supporting details. Comprehension derives not only from situational and subject-matter knowledge but also from knowledge of the language itself.

Level 3

Evaluative mode.

Author purpose: Evaluate situations, concepts, conflicting ideas; present and support arguments and/or hypotheses with both factual and abstract reasoning; often accompanied by the appropriate use of wit, sarcasm, or emotionally-laden lexical choices.

ILR descriptor: Text-types include news stories similar to wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field; all of these may include hypothesis, argumentation, and supported opinion.

ACTFL version: Texts that use precise, often specialized vocabulary and complex grammatical structures. These texts feature argumentation, supported opinion, and hypothesis, and use abstract linguistic formulations as encountered in academic and professional reading. Such texts are typically reasoned and/or analytic and may frequently contain cultural references.

PART TWO: SAMPLES OF TEXTS AT LEVELS 1, 2, AND 3

Level 1

An announcement made in a newspaper.

SUNDAYS

Dance at the American Legion Hall.

3-6 p.m. the second and fourth Sunday of the month at Dolores and Eighth, Carmel.

Dance to 1940s and 1950s music by local bands.

Level 2

A factual news report.

Overnight closures of lanes and ramps on and leading to the Bay Bridge will conclude this morning, except for the Essex Street on-ramp, which will remain closed until 2 p.m.

At 6 a.m., the First Street on-ramp will reopen along with two eastbound lanes on the bridge. Three westbound lanes on the bridge will reopen at 5 a.m.

All closures are subject to change. For updated information, visit www.baybridgeinfo.org or www.511.org.

Level 3

A newspaper editorial.

SOMETHING IN HONG KONG'S AIR

Anyone who has ever fallen under Hong Kong's exotic spell in recent years knows how taking a deep breath can make the magic disappear. The air throughout Hong Kong's tropical landscape has grown steadily more polluted—tainted by dark, unhealthy clouds from power plants, traffic and under-regulated smokestacks from the Chinese mainland.

Hong Kong's average air pollution levels can be so high—double or even triple the World Health Organization limit—that some analysts estimate the air contributes to an extra 2,000 deaths a year. Leaders in Beijing and Hong Kong have repeatedly promised to cut down on environmental toxins in the air, land and water.

But with China's booming economy, such promises keep sliding down the real priority lists. What might change that attitude is how the outside business world views the quality of life for employees. As one businessman explained to The Wall Street Journal about his family's recent retreat to Australia: "You can drink bottled water. But with the air –you have to breathe it."

Such departures have finally begun to raise concerns in Hong Kong's business community. The local Chamber of Commerce issued an urgent request for the government to commit to "genuine reductions in air pollution" after it found that "an alarming 95 percent" of executives interviewed were worried or very worried about air quality and its effects on their health. But in a disheartening development this week, Hong Kong's chief executive, Donald Tsang, missed yet another opportunity to lay out a workable plan for clearing the air quickly.

This is not a hopeless situation, as leaders in Mexico City could attest. Once a place where residents courted asthma with every step outside, Mexico City approved what is generally regarded as one of the best and most comprehensive approaches to air pollution in 1990. The measures included everything from new fuel composition standards to new emission standards for vehicles. As a result, Mexico City halved some forms of air pollution in only five years. If Hong Kong even committed to cutting its pollution in half, that would be a good start.

PART THREE: WHAT THE PROFICIENCY SCALES SAY ABOUT TASKS CHARACTERISSTIC OF EACH DIFFICULTY LEVEL

Level 1

Orientational mode.

Reader purpose: Orient oneself by identifying topics and main topic or fact(s).

ILR descriptor: Can get some main ideas and identify general subject matter in some authentic texts.

ACTFL version: Able to understand messages found in highly familiar, everyday contexts.

Level 2

Instructional mode.

Reader purpose: Understands not only the main topics and facts, but also the supporting details, such as temporal and causal relationships.

ILR descriptor: Can locate and understand the main ideas and details; able to answer factual questions about Level 2 texts.

ACTFL version: Can understand the main ideas, facts, and many supporting details.

Level 3

Evaluative mode.

Reader purpose: Learn by relating ideas and conceptual arguments. Understand the text's literal and figurative meaning by reading both "the lines" and "between the lines." Recognize the author's tone and infer the author's intent.

ILR descriptor: Can understand hypothesis, argumentation, and supported opinion. Able to interpret material correctly, relate ideas, and "read between the lines."

ACTFL version: Can draw inferences from textual and extralinguistic clues. Can understand texts that feature argumentation, supported opinion, and hypothesis, and use abstract linguist formulations as encountered in academic and professional reading.

PART FOUR: SAMPLES OF MULTIPLE-CHOICE TEST ITEMS FOR EACH OF THE SAMPLES SHOWN IN PART TWO.

Note:

- The first three items apply tasks to texts of the same level.
- The next items show the special case of writing Level 1 items for higher level texts.

Level 1

A newspaper announcement:

What is announced?

A. Dancing *
B. Music lessons
C. A history class
D. A picnic

Task: The reader is asked to identify the topic of the text.

The item is intended to discriminate between solid Level 1 readers and those below that level.

Level 2

A news report:

What is reported?

A. Times that lanes and ramps will re-open.*
B. An accident on Essex Street at 2 p.m.
C. Construction of a new on-ramp to the bridge.
D. Heavy traffic on two east-bound lanes

Task: The reader is asked to locate the main idea and details of the text.

The item is intended to discriminate between solid Level 2 readers and those below that level.

Level 3

From the editorial page:

What does the writer imply about China's priorities?

A. Financial considerations are the major factor in setting policy.*
B. China is determined to meet its targets to attract foreign businesses.
C. Cultural differences would prevent their using Mexico's approach.
D. A commitment to cutting air pollution in half is still on the agenda.

Task: The reader is required to relate ideas and draw inferences.

The item is intended to discriminate between solid Level 3 readers and those below that level.

SPECIAL CASE FOR LEVEL 1

Sometimes we have been required to test Level 1 reading skills when no Level 1 texts were available.
We have used items like the following for that purpose:

Level 2 text

A news report:

According to the report, when will this take place?

A.Today *
B. Tomorrow
C. Tonight
D. In two days

Level 3 text

From the editorial page:

The number 2,000 refers to

A. people who die each year. *
B. departures for Australia.
C. new businesses.
D. power plants.

PART FIVE: THE ITEM WRITING PROCESS

Level 2 Text

1. Select a factual, concrete text.

Bradley Manning didn't complain about mistreatment, prosecutors contend

Prosecutors try to counter Bradley Manning's claims of abuse in confinement.

The hearing focuses on Manning's time in the military brig at Quantico, Virginia.

Defense wants case dismissed on grounds that Manning's confinement was harsh.

The Army private is accused of stealing thousands of classified documents.

Prosecutors tried to establish Friday that Army private Bradley Manning—charged

in the largest leak of classified material in U. S. history—missed multiple

opportunities to complain about the mistreatment he's alleging he suffered in military

custody.

While cross-examining Manning at a pre-trial hearing at Ft. Meade, Maryland, prosecutor Maj. Ashden Fein asserted that records of weekly visits Manning had with unit officers during nine months of detention at Quantico, Virginia, show no complaints about his treatment.

The cross-examination—during a hearing on a defense motion to have Manning's case dismissed on grounds that his confinement has been harsh and has amounted to enough punishment—came a day after Manning testified that he had considered suicide while in custody.

The Army intelligence analyst, arrested in June 2010, is accused of stealing thousands of classified documents while serving in Iraq.

The material was then published online by WikiLeaks.

WikiLeaks has never confirmed that Manning was the source of its information.

2. Review the appropriate tasks for this level. These include locating the main idea and details as well as answering factual questions about the task.

3. Draft a key (correct answer) requiring examinees to perform Level 2 tasks.

Possible keys for this text
Include:

- Manning complained he was mistreated while confined.
- Prosecutors countered Manning's charges of abuse.
- Thousands of classified documents were leaked.

4. Experiment with distractors (incorrect responses). Combined with a stem (or lead-in), the key and distractors form a set of options. Draft a stem and distractors to fit with one of these keys.
Possibilities are:

What issue was raised by the defense?

- Manning's suicide attempt shows the brig was mismanaged.
- Manning has already served more jail time than other leakers.

OR

According to prosecutors,

- Manning missed appointments to discuss prison conditions.
- unit officers inspected Manning's cell weekly for nine months.
- additional classified documents were published after Manning was jailed.

5. Draft an item with the stem and options you prefer. One possible item:

What does this report state about Bradley Manning's case?

A. He complained he was mistreated while confined. *
B. He attempted suicide because of brig conditions.
C. Unit officers inspected his cell weekly for nine months
D. Officials said he missed appointments to discuss abuse.

6. Review the item.

- Is there a correct response? Yes, Option A is correct according to the text.
- Is there only one correct response? Yes.

The distractors are wrong because:

- Option B---Manning considered but did not attempt suicide.
- Option C---Manning had weekly visits with unit officers; nothing is said about inspections of his cell.
- Option D---Prosecutors said he missed opportunities to claim abuse, not appointments.

*Are the distractors plausible? Yes, all fit within the subject matter of the text.

*Would many examinees with proficiency below Level 2 be likely to select them? Yes, lower level examinees trying to match words they recognize would find language from the text in these distractors:

Option B—Contains “suicide” and “brig.”
Option C—Contains “unit officers,” “weekly,” and “nine months.”
Option D—Contains “missed” and “abuse.”

*Do any options stand out as different from the others? No option is particularly longer or shorter than the others. None is constructed in an unusual way.

Level 3 Text

1. Select an evaluative text presenting an argument that includes abstract concepts.

In a Constantly Plugged-In World, It’s Not All Bad to Be Bored

I spent five unexpected hours in an airport this Thanksgiving holiday when our plane had mechanical difficulties and we had to wait for another plane to arrive.

So I had plenty of time to think about the subject of boredom.

I won’t lie to you.

Half a day in an airport waiting for a flight is pretty tedious, even with the distractions of books, magazines and iPhones (not to mention duty-free shopping).

But increasingly, some academics and child development experts are coming out in praise of boredom.

It’s all right for us—and our children—to be bored on occasions, they say.

It forces the brain to go on interesting tangents, perhaps fostering creativity.

And because most of us are almost consistently plugged into one screen or another these days, we don’t experience the benefits of boredom.

So should we embrace boredom?

Yes.

And no.

But I’ll get back to that.

First of all, like many people, I assumed that boredom was a relatively recent phenomenon, with the advent of more leisure time.

Not so, says Peter Toohey, a professor of Greek and Roman history at the University of Calgary in Canada and the author of “Boredom: A Lively History” (Yale University Press, 2011).

“Boredom actually has a very long history,” he said.

There’s Latin graffiti about boredom on the walls of Pompeii dating from the first century.

Then there’s the question of how we define boredom.

The trouble is that it has been defined, and discussed, in many different ways, said John D. Eastwood, an

associate professor of psychology at York University in Ontario, Canada.

After looking over the research literature and putting the idea in front of a focus group of about 100 people, Professor Eastwood and his colleagues defined boredom as an experience of "wanting to, but being unable to engage in satisfying activity."

What separates boredom from apathy, he said, is that the person is not engaged but wants to be.

With apathy, he said, there is no urge to do something.

The core experience of boredom, he said, is "disruption of the attention process, associated with a low mood and a sense that time is passing slowly."

Boredom can sound an awful lot like depression.

But Professor Eastwood said that while they can be related, people who are bored tend to see the problem as the environment or the world, while people who are depressed see the problem as themselves.

Sometimes we think we're bored when we just have difficulty concentrating.

In their study, "The Unengaged Mind: Defining Boredom in Terms of Attention," which appeared in the journal Perspectives on Psychological Science in September, Professor Eastwood and his colleagues pointed to an earlier experiment in which participants listened to a tape of a person reading a magazine article.

Some groups heard a loud and unrelated television program in the next room, others heard it at a low level so it was barely noticeable, while the third group didn't hear the soundtrack at all.

The ones who heard the low-level TV reported more boredom than the other two groups—they had difficulty concentrating but were not sure why, and attributed that difficulty to boredom.

When you're trying to focus on a difficult or engaging task, disruption of attention can lead to boredom, said Mark J. Fenske, an associate professor of neuroscience at the University of Guelph in Ontario and one of the authors of the study.

On the other hand, when you're doing something dull, "such as looking for bad widgets on a factory line, distracting music can help you not be bored."

In fact, he said, we now know that squirming and doodling, often seen as a sign of boredom, can actually help combat it by keeping people more physically alert.

"Research shows that kids who are allowed to fidget learn more and retain more information than those who are forced to sit still," Professor Fenske said.

2. Review the appropriate tasks for this level. These include relating ideas and conceptual arguments, understanding hypothesis, argumentation, and supported opinion, and reading "between the lines."

3. Draft a key (correct answer) requiring examinees to perform Level 3 tasks. A possible stem (lead-in) and keys for this text include:

According to the research cited,

*environmental interference with an interesting activity can cause boredom.
*the brain's activities during periods of boredom may enhance creativity.

4. Experiment with distractors (incorrect responses). Possibilities are:

*apathetic people get bored because they have do not have an urge to be active.
*distractions such as television are more likely to cause depression than boredom.
*although boredom has a long history, its definition has changed frequently.
*a consistent ability to resist boredom is essential for fostering creativity.

5. Draft an item with the stem and options you prefer. One possible item:

According to the research cited,

A. environmental interference with an interesting activity can cause boredom.*

B. apathetic people get bored because they do not have an urge to be active.
C. a consistent ability to resist boredom is essential for fostering creativity.
D. distractions such as television are more likely to cause depression than boredom.

6. Review the item.

*Is there a correct response? Yes, Option A is correct according to the text.

*Is there only one correct response? Yes. The distractors are wrong because:
Option B.—The text only mentions apathy to contrast it with boredom.
Option C.—The text states that boredom may enhance creativity.
Option D.—While the text mentions television and depression, it does not link them.

*Are the distractors plausible? For the non-expert in this field, the distractors should be plausible.

*Would many examinees with proficiency below Level 3 be likely to select them? A Level 2 reader will probably try to connect ideas and language from the text but lack the analytical ability in the language to draw the correct conclusions.
Option B.—The text distinguishes apathy from boredom (in the opinion of one researcher). A lower level reader may not realize the text does not state whether apathetic people can also experience boredom.
Option C.—The text says some experts think boredom can foster creativity. However, a lower level reader may miss that counterintuitive point and accept a negative statement about boredom.
Option D.—The text contrasts depression and boredom and also mentions an experiment with television sound. A lower level reader may be drawn to an option using these words.

*Do any options stand out as different from the others? The options are all about the same length. (It would not be a good idea to have three longer sentences and one short phrase, for example.) All refer explicitly to “boredom” or “bored.” All fit with the stem grammatically.

PART SIX: 45 MINUTE WORKSHOP

We will work with these three documents:

- Doc-A primerahora/2012/12/01/127594 (Segments 1-22)
- Doc-B derstandart.at/2012/12/01/141907 (Segments 1-20)
- Doc-C dw/2012/12/01/82217 (Segments 1-30)

0:00 – 0:05 Introductory remarks about ILR, distribute instructions. Workshop participants will be assigned in equal numbers to 9 different teams.

0:05 – 0:30 Each team collectively writes one test item according to team assignment, using the instructions and background in Parts 1-5 of this document.

- Team 1: Write Level 1 Question for Doc-A reference translation
- Team 2: Write Level 1 Question for Doc-B reference translation
- Team 3: Write Level 1 Question for Doc-C reference translation
- Team 4: Write Level 2 Question for Doc-A reference translation
- Team 5: Write Level 2 Question for Doc-B reference translation
- Team 6: Write Level 2 Question for Doc-C reference translation
- Team 7: Write Level 3 Question For Doc-A reference translation
- Team 8: Write Level 3 Question For Doc-B reference translation
- Team 9: Write Level 3 Question For Doc-C reference translation

(Each question will be written on a poster board or projected for everyone to see)

0:30 – 0:40 Each participant individually answers three questions for two documents each (total of six questions) according to these assignments. Allows 5 minutes per document, which is a typical time allotment.

- Team 1 Members: Doc-B in MT-1 Doc-C in MT-2
- Team 2 Members: Doc-A in MT-3 Doc-C in MT-1
- Team 3 Members: Doc-A in MT-2 Doc-B in MT-3
- Team 4 Members: Doc-B in MT-2 Doc-C in MT-3
- Team 5 Members: Doc-A in MT-1 Doc-C in MT-2
- Team 6 Members: Doc-A in MT-3 Doc-B in MT-1
- Team 7 Members: Doc-B in MT-3 Doc-C in MT-1
- Team 8 Members: Doc-A in MT-2 Doc-C in MT-3
- Team 9 Members: Doc-A in MT-1 Doc-B in MT-2

(Note that participants do not answer questions for the document they wrote questions for since they were already exposed to it)

0:40 – 0:45 Correct answers shown to participants; participants score answers and turn in tests for tabulation and circulation of results to participants.

CPT-MT WORKSHOP COMPREHENSION TEST

Participant name:	
Email address:	
	Circle one on each row for Team Number and Doc Conditions
Team Number:	1 2 3 4 5 6 7 8 9
Doc A Condition:	MT-1 MT-2 MT-3 Question-Writing
Doc B Condition:	MT-1 MT-2 MT-3 Question-Writing
Doc C Condition:	MT-1 MT-2 MT-3 Question-Writing

Question	Answer (A, B, C or D) or NA (Not Applicable)	Correct (1=yes; 0=no) (leave blank if NA)
1. Doc-A Question 1 (Level 1)		
2. Doc-A Question 2 (Level 2)		
3. Doc-A Question 3 (Level 3)		
4. Doc-B Question 1 (Level 1)		
5. Doc-B Question 2 (Level 2)		
6. Doc-B Question 3 (Level 3)		
7. Doc-C Question 1 (Level 1)		
8. Doc-C Question 2 (Level 2)		
9. Doc-C Question 3 (Level 3)		

COMMENTS ON SELECTED DOCUMENTS

Doc-A: primerahora/2012/12/01/127594 (Segments 1-22)

- This portion of text 127594 provides a brief analysis of a type of course popular on campus in 2012.
- Although it is largely factual, the text contains some abstract language and concepts, as well as colloquialisms (e.g., "at 30 days and counting") not usually found in print. The author's point of view is reflected only in his/her choice of examples to cite. The proposed test item is intended to identify a statement that generally reflects the information present in the text.
- This text and item might distinguish between those machine translation systems that do and do not include all essential factual information in the original in a grammatically accurate form.
  This portion of the text would probably be rated as a very low Level 3 or even Level 2+.

Doc-B: derstandart.at/2012/12/01/141907 (Segments 1-20)

- This portion of text 141907 is evaluative in that the author demonstrates a definite point of view and colors factual information with humorous language. It has the light-hearted tone of a newspaper feature story, not a serious report. The proposed test item is intended to identify basic factual information embedded in this humorous text.
- When used to evaluate the quality of machine translation systems, this text and item could distinguish between systems that do and do not present key factual points in the text. It would be important that the system preserve the grammatical accuracy of the original.
- Unless researchers were interested to measuring the system's capacity for conveying an arch or sarcastic tone, this would not be a good text for evaluating an MT system. Testing tone (the author's attitude toward the subject) adequately would probably require that examinees be tested individually and asked to paraphrase the text or write an essay explaining how the choice of words throughout the text conveyed the attitude.
- Doc-C dw/2012/12/01/82217

Doc-C: derstandart.at/2012/12/01/141907 (Segments 1-30)

- This portion of text 82217 is an example of an argument that analyzes information, explores some abstract concepts, and draws conclusions from the facts presented. It reflects the opinion of one person—Stephen Szabo. The proposed test item is intended to identify a major point made by Szabo in the text. It requires a certain amount of synthesizing information.
- When used to evaluate the quality of machine translation systems, this text and item could distinguish between systems that do and do not include all major points in the text, including those dependent on abstract language, with grammatical accuracy that supports the meaning of the original.
- The text is approximately Level 3. A simple main idea item would not be appropriate.