# Language Skill Level Descriptions

| Level 0+ | Level 1 + | Level 2 + | Level 3 + | Level 4 + | Level 5 |
|---|---|---|---|---|---|
| ● | ● | ● | ● | ● | ● |
| **Listening**<br>understands certain memorized utterances in areas of immediate needs with extralinguistic cues | **Listening**<br>understands basic survival utterances, simple questions and answers on familiar topics main ideas | **Listening**<br>understands routine conversations and discourse about familiar topics; gleans all the facts | **Listening**<br>understands essentials of all speech; grasps opinion and inferences | **Listening**<br>understands all forms and styles of speech, even some non-standard dialects; develops and analyzes argumentation | **Listening**<br>understands extremely difficult and abstract speech and how natives think as they create discourse |
| **Reading**<br>reads alphabet or high-frequency characters, recognizes some numbers and isolated words | **Reading**<br>reads simple, predictable material in print or type; identifies general topics | **Reading**<br>reads simple, authentic, straightforward material on familiar topics; uses contextual cues | **Reading**<br>reads a variety of prose on unfamiliar subjects that may include opinions, hypothesis and analysis | **Reading**<br>reads fluently and accurately all styles and forms; grasps full ramifications of texts within wider context | **Reading**<br>reads very difficult and abstract prose |
| **Speaking**<br>produces telegraphic utterances for immediate survival needs | **Speaking**<br>maintains very simple conversations on familiar topics; cannot produce continuous discourse unless rehearsed | **Speaking**<br>handles routine, high-frequency, limited interactions and conversations about current events, family and common topics | **Speaking**<br>participates effectively in most formal and informal conversations about practical, social, and professional topics within a shared context | **Speaking**<br>uses the language fluently and accurately for all purposes | **Speaking**<br>commands language with complete flexibility and intuition; pronunciation consistent with that of educated native speaker |

# Machine Translation Evaluation Methodology

**Jim Baker, Peter Brown, Lynn Carlson, Eduard Hovy, Charles Wayne, John White**

**June 24, 1992**

## 1.Introduction

This document outlines the methodology for evaluating translation systems in the DARPA machine-translation initiative. Evaluations will measure both the savings in human time realized by using a computer as a translation aid, and also the quality of resultant translations. There is a trade-off between the human effort required to produce a translation and the quality of that translation. This trade-off will be explored by plotting normalized human translation time (see Section 6 below) against measures of translation quality.

Each system will be evaluated on the translation of documents from some source language, which may be different for different systems, into English. System builders must choose to have their system evaluated either as a stand-alone translation system or as an aid to human translation. At their discretion, system builders may choose to have their system evaluated in both of these roles. Only Level II translators will be used when a system is being evaluated as an aid to human translation.

Translations will be judged for accuracy and style by a panel of experts who will rate each sentence on an 8-point system. In addition, monolingual English speaking subjects will take multiple choice comprehension tests on passages translated into English, and the semantic fidelity of the translated passages will be evaluated by scoring the subjects' responses.

## 2.Passages

Each passage used in an evaluation will be a news article, or a portion thereof, on the subject of mergers and acquisitions. Each passage will be between 300 and 500 words in length. For each language pair, the passages used in an evaluation will consist of one set of passages originally produced in the source language (for example, passages from Le Monde if the source language is French) and another set of equally many passages translated from English into the source language (for example, passages from the New York Times translated into French). The English passages from which the passages in this latter set are derived are referred to as master passages. In an evaluation, the same set of master passages will be used for all language pairs. The original language passages are included in order that evaluations can be made on naturally occurring source text. Translations of the master passages are included in order to minimize differences in the difficulty of passages due to differences in their semantic content. These translations are produced by highly skilled expert translators.

A document describing the format of the passages, as well sample passages in this format, will be provided to system developers at least one month prior to the start of an evaluation. These sample passages will include at least two passages translated from master passages as well as two original source-language passages. The passages in this sample will be used by system developers to prepare their systems for evaluation and also to calibrate human translators. No passage in this sample will be used in the evaluation, or in any later evaluation.

## 3. Rules of the Game

All passages used in the course of an evaluation will be kept secret from system developers until the evaluation begins. Before evaluation passages are revealed, the programs and databases used by all systems being evaluated must be frozen, and may not be modified until after the evaluation is complete. In case a system must be modified in order to run at all, full documentation of all modifications must be provided to test administrators. It is, of course, permissible for a system to automatically adjust stored internal parameters as it runs and/or for translators using a system to add facts (e.g. new words) to the system's database (at the expense of translation time). Errors or formatting problems in the test passages may be corrected only by test administrators. Pre-editing of test passages or post-editing of translations may be performed only by translators officially participating in the evaluation; all such activities must be carefully timed and shall constitute a part of the human time used in making the translations.

## 4. 8-point Evaluation of Accuracy and Style

A panel of bilingual experts will be used to evaluate the quality of the various versions of a translated document. The methodology is based on a system which has been developed and used within the U.S. Government for evaluating human translators. This methodology is externally motivated, and represents a clearly established and accepted standard within a particular community.

Each version of a translated document is compared to the source language original and evaluated on a per sentence basis, taking into consideration the context of the overall document. Within a sentence errors are classified according to a three-way distinction, and weighted accordingly. Syntactic errors which result in a corresponding semantic error are assessed a four point deduction. These include 1) incorrect assignment of case roles, resulting from incorrect assignment of the subject or object, for example, 2) interclausal errors, such as misidentification or misplacement of relative or subordinate clauses, and 3) attachment errors involving prepositional phrase modifiers. Lexical errors are assessed a two-point deduction, and include both incorrect vocabulary items as well as morphological errors. Errors in English style or usage are assessed a one-point deduction. A maximum of eight points can be deducted for any given sentence. Once the eight-point cap is reached, the evaluator moves on to the next sentence without further analysis.

## 5. Comprehension Test

In an evaluation, each master passage gives rise to a family of English passages consisting of the passage itself, and all of its round-trip translations.

In these round-trip translations we include English translations by humans, by machines, and by human-aided machines. Suppose, for example, that two systems are being evaluated, one with French as the source language, and one with Russian. Suppose further that the system builders for each system choose to have their system evaluated both as a stand alone system and as an aid to human translation, Then each master passage will have a family of seven translations: a human translation into English of its translation into French, the machine-alone translation into English of its translation into French, the human-aided machine translation into English of its translation into French, a human translation into English of its translation into Russian, the machine-alone translation into English of its translation into Russian, and the human-aided machine translation into English of its translation into Russian, and finally the original master passage itself (not technically a translation). Define each member of a family as a version. So, for example, a human-aided machine translation from French by System A might be one version, a human-aided machine translation from French by System B might be another version, and a human translation from Russian might be yet another version. Although the number of members in a family will depend on the choices of system builders as to whether their systems will be evaluated in one way or in two ways, the family for each master passage will have the same number of members. Let this number be N. In order to have a balanced test, the number, T, of test takers used in an evaluation must be a multiple of N. Furthermore, the number of master English passages must be a multiple of T. Comprehension tests will then be administered so that each test taker is tested on exactly one member from each family, and furthermore so that each test taker is tested on on each version the same number of times. During the course of the comprehension tests, test takers will only be exposed to English texts, not to any of the foreign-language source text.

For each master passage, a set of four, five, or six multiple choice questions will be constructed. For each question, there will be a set of five possible answers determined as follows. The testing organization will construct a correct response and four incorrect responses. The incorrect responses will be placed in positions (1) through (4), and the response 'none of the above' will be placed in position (5). A 6-sided die will then be rolled until a number other than 6 appears on the top face. If the die roll is 5, the responses already placed will be unaltered. Otherwise, if the die roll is R, the response in the Rth position will be discarded and replaced by the the correct response. In this way, each question shall have exactly one correct response and the probability of any position containing the correct response shall be 1/5.

Care should be taken when constructing questions and answers to be used in a comprehension test. It should not be the case that certain answers can be ruled out without information from the passage pertaining to the question. Furthermore, questions should not give away answers to other questions. To a limited extent this can be controlled for by asking test takers to attempt to answer questions without access to any version of the relevant passages.

## 6. Normalization of Human Effort

All of the human translators participating in an evaluation will be Level II translators, according to the language skill levels established by the Interagency Language Roundtable and adopted government-wide by the Office of Personnel Management in 1985. Nevertheless, there still will be differences in their skills that may be reflected in differences in the speed with which they are able to produce translations. There may also be inherent differences in the ease of translation from one language pair to another that will affect the speed with which translators are able to produce translations. In order to compensate partially for these differences, each translator will translate all sample passages in his source language prior to an evaluation. He will be asked to translate to the same level of quality as he is asked to translate to during the evaluation, and the time he requires to translate each passage will be recorded. These times will be averaged, and will serve as a basis against which to compare his performance during the evaluation. A normalization factor will be computed for each translator by dividing his average time on the sample passages by the average of the average times for all translators. When plotting results of human-aided machine translations, all times measuring amount of human effort will be divided by the appropriate normalization factors.

## 7. Reporting Results

For each version (as defined in Section 5) means and variances will be computed for normalized human-effort times and quality scores for all original language passages and separately for all master passages. Means and variances for scores on comprehension tests will be computed from measurements on all master passages.

Four x-y charts will be produced from the results of an evaluation: 1) time vs. quality on original language passages; 2) time vs. quality on master passages; 3) time vs. quality on both types (master and original language) of passages; and 4) time vs. comprehension score on master passages. In each chart, the x-axis will denote normalized time, and the y-axis will denote either quality or comprehension score. All versions will be plotted on each chart. On the fourth chart, the results from comprehension tests on the original master passages will also be plotted. The results for each version will be plotted as a rectangle, two standard deviations by two standard deviations, centered at the time-mean and score-mean.

## 8. Logs

During the course of an evaluation, those humans interacting with systems may make various observations about those systems that will be of use to system developers. Administrators of an evaluation will keep a log of any such observations and may choose to elicit more observations through questionnaires. Alter the evaluation is complete any such logs and completed questionnaires will be provided to the appropriate system developers.