Direct Application of A Language Learner Test To MT Evaluation

Florence Reeder
The MITRE Corporation
7515 Colshire Dr.
McLean VA 22102
freeder@mitre.org

Abstract

This paper shows the applicability of language testing techniques to machine translation (MT) evaluation through one of a set of related experiments. One straightforward experiment is to use language testing exams and scoring on MT output with little or no adaptation. This paper describes one such experiment, the first in a set. After an initial test (Vanni & Reeder, 2000), we expanded the experiment to include multiple raters and a more detailed analysis of the surprising results. Namely that unlike with humans, MT systems perform more poorly at both level zero and one than at level two and three. This paper presents these results as an illustration of both the applicability of language testing techniques and also the caution that needs to be applied.

1 Introduction

One basic test that can be performed is the direct application of language learning tests to MT system outputs. The impetus behind this experiment comes from funding stakeholders, particularly in the United States government. In an effort to better understand an unfamiliar technology and to quantify what MT can provide, these stakeholders try to frame system evaluations in terms that are familiar to them, specifically, human translator ratings. For such a community, a key question would therefore be: Can scoring criteria for establishing language learner ratings be utilized, *as is*, for scoring MT output? This experiment was designed to contain little or no adaptation for MT. The resulting data set gives indicators of the differences between evaluating language learners and machine translation.

The scale used is an Interagency Language Roundtable (ILR) style scale (Lowe & Stansfield, 1988). This is theoretically attractive, as the scale results from years of pedagogical research in understanding and measuring student proficiency in foreign languages. Also, the scoring system is well-known among the intended evaluation consumers. The proficiency scale is task-based, so the evaluation can address the needs of multiple stakeholders by answering questions about the functions a student can perform, some of which parallel the uses of MT. From this ILR scale comes a set of teaching recommendations. Because of the detailed characteristics described and the test items designed for each level, scoring with this scale permits users to examine relevant linguistic features that each system provides. Test items are from a ready-made corpus that has been vetted through years of test administration. Finally, since these guidelines and materials are currently being used by the US government, a pool of trained raters exists for multiple languages.

2 Experimental Setup

For this experiment, five texts were selected, one for each of the first five ILR levels. The texts were run through three MT systems. Four judges looked at each text in order of increasing difficulty. The sample texts were taken from a course based on the ILR scale. The language pair in this experiment is French-English.

The test items are based on the ILR scale and associated guidelines for evaluating students to an ILR level. The scale is designed to show gradations of ability. Instructors consider level two to be the lowest proficiency level at which language competence can be claimed (Lowe & Stansfield, 1988). The test items used in this experiment reflect an emphasis on translation as a required skill for the given language school’s graduates and as such are drawn from realistic data sets. The texts

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 166-175,
Cambridge, August 2006. ©2006 The Association for Machine Translation in the Americas

come from news items used for testing at the school and are characteristic of a given level.

Test items in level zero (T0) are characterized by very simple phrases, idioms, single words and other basic vocabulary knowledge elements. Items in this category include lists of newspaper feature headings and represent basic survival knowledge. Grammatical constructions are simple, primarily noun phrases with noun-noun modifiers, noun conjunctions and single preposition noun phrases.

Basic, simple sentences characterize level one (T1) test texts. The data is grammatical, but does not represent the full range of grammatical utterances. It is idiomatic at the level of a tourist phrase book. A test at this level would address basic conversational skills with minimal personal information. The data consists of short sentences, with proper names such as **Jules**, **Jaques**, times or place names, and is more conversational in style than typical written material. The percentage of closed class words at this level is significantly higher than in the previous level (Table 1)[1].

Table 1: Word count and percentage closed class words

LEVEL	WORD COUNT	% CLOSED CLASS
T0	125	16.0
T1	164	56.1
T2	155	43.2
T3	128	49.2
T4	218	56.4

Level two texts (T2) have sentences with complex grammatical constructions, but without complex semantic features. They are taken from news texts and represent coherent language usage about real-world topics. Syntactically they incorporate modifying clauses and phrases. The percentage of proper nouns is increased, as is the number of specialized terms. The percentage of closed class words such as prepositions, articles and conjunctions is roughly half of the words (Table 1).

Level three (T3), like level two, is characterized by complex sentences. At level three, complex complement structures and sophisticated attachments are used to express opinion and give some depth of perspective. The distinction between input texts at level three versus two is subtle, dealing with less concrete events. For example, rather than reporting on an official's state visit, they report on policy changes within the government.

Level four (T4) contains flowing language with complex idiom, inference and subtle usages intermixing freely. Pieces at this level include editorial pieces with analogy and illustration. These thought pieces represent fluency in a language, although not quite at the well-educated, native speaker level.

2.1 Test Data

In this experiment, the data was selected by a collaborator (Vanni & Reeder, 2000) who taught French for translators under the ILR guidelines. The teaching and testing materials were extracted primarily from French language newspapers. Selections were made appropriate to each level of language ability, zero through four, from texts used in a previous semester's classes. While the length of a typical test text is approximately 600 words, the sample texts used here were shortened to between 125 and 225 words due to constraints on the initial iteration of the experiment. Since scoring is normalized on the target language word count in each text, the shortening should have little effect on the overall scores. Expert translations of each text were produced by the teacher. An example level one text is shown (Figure 1).

```
Bonjour, comment allez-vous?
Ça va bien, merci.  Et vous?
Tout va bien.
Bonne journée.
Comment vous appellez-vous?
Je m'appelle Jeanne, et eux?
Ils s'appellent Jacques et Jules.
```

```
Hello, how are you ?
Very well, thank you and you ?
Everything is going well.
Have a nice day.
What's your name ?
My name is Jeanne.  And them ?
Their names are Jacques and Jules.
```

Figure 1: L1 Text: Source and Human Translation

3 Experiment Execution

Three MT systems of varying degrees of sophistication and development were used for the initial

[1] Closed class words were counted according to a standard list of closed class words from an information retrieval project.

experiment[2]. System S1 represents a lexical transfer system with minimal development, but with some lexical specialization for the types of data typically seen in the student’s environment. System S2 is a grammatical transfer system where development and specialization had ceased approximately three years before the experiment. System S3 is represented by the vendor as the most sophisticated available, with both semantic and syntactic processing. The system continues to be expanded and updated, after more than ten years of development.

The system outputs were converted into a uniform format, primarily through the process of reentering printed outputs. There were two reasons for doing this. First, any system-identifying characteristics or tags needed to be removed. Second, the texts were not readily available in electronic form[3]. While it could be argued that some systems would benefit from being re-keyed, every effort was made to preserve all relevant features of the output text. That is, the punctuation, capitalization and unit boundaries were entered exactly as they appeared on the printouts, whereas system tags such as **$$LN**[4] were removed. Since there were no judgments as to the “humanness” of the data and since all participants knew they were grading MT output, no attempt to correct or normalize the data was made beyond removing system tags. For instance, system S1 marks all untranslated words in upper case, a convention that was preserved. Another system put a header identifying it in the text and this was removed. Samples corresponding to the text in Figure 1 are included here (Figure 2).

Table 2: Grading Scheme

Error	**Description**	**Score**
SYN	Major syntactic errors significantly altering meaning. Examples are subject as object, wrong modifier attachment	**-4**
LEX	Lexical errors (tense, number, agreement, omitted words). In this case, not translated words belong here.	**-2**
AWK	Stilted usage; disfluencies	**-1**
SLO	Difficult but understandable; punctuation error	**-0.5**

After they were converted to a uniform format, the texts were given to language teachers and professionals for scoring. In the initial experiment, there was only one scorer (Vanni & Reeder, 2000). Three additional scorers are presented here. The scorers received scoring directions according to the Professional Qualifying Exam guidelines used for scoring an ILR-based scale. In this scoring, errors are assigned and marked (Table 2).

In addition to getting a description of the types of errors which fall into each category and examples of these, the scorers were given instructions for situations such as multiple source errors. In cases where an error had multiple possible causes or where there were multiple levels of error in a given location, graders were instructed to judge the most serious cause and base their assessment only on that. Therefore, if there is an agreement error due to an untranslated word, a single marking of –2 would be assigned for the lexical error. This puts a burden on the graders to attribute error causes, something that research has shown to be a difficult task (e.g., Schwind, 1994; Holland & Kaplan, 1995; Heift, 1998; Michaud & McCoy, 1999).

Figure 2: Level 1 Translation Sample for Systems

SYSTEM	**SAMPLE TEXT**
S1	hello, how do you do ? it’s going well, thanks. and you ? all is going well. good day. how you called-you ? my name is JEANNE, and them ? they ‘called JACQUES and JULES.
S2	hello, how are you ? That is fine, thank you. And you ? Whole is fine. Good day. how You appellez – you ? I Am called Jeanne , and them ? They are called Jacques and Jules.
S3	Hello , how are you ? That is well , thank you . And you ? All is well . Good day How do you appellez yourselves ? I am called Jeanne , and them ? They are called Jacques and Jules

[2] Due to anonymity requests, the particular systems cannot be named. One is a US Government system (S1) and two are commercially available (S2, S3).
[3] The systems were available only on an internal network at the time of the experiment which prevented the dissemination of the output in electronic form.
[4] System S3 used this to delineate line breaks.

A final score is arrived at by counting the number in each error category, multiplying this by the error weight, adding the scores for all the categories and then subtracting this from the number of words. Because of the variations between language learners and MT systems, some adaptation of the test methodology was necessary even for this initial experiment. The first difference is the decreased length of the texts. Instead of using a 600 word text, the texts ranged from 126 to 219 words. Scores were adjusted accordingly, by normalizing for word count. The second difference is an adaptation of scoring instructions. In scoring students, the purpose is to ascertain whether or not the student has attained a given level of competence. Therefore, the scoring is pass-fail and the scorer stops when she has subtracted more than twice the number of points in the pass-fail cut-off (5% of target words). In this experiment, the interest is in relative system ranking, scorers were instructed to grade the entire text, whether or not the cut-off was reached. The final adaptation is related to the attribution of errors. Research has shown that MT systems make some kinds of mistakes that human translators or language learners would rarely make (Lowe & Stansfield, 1988; Schwind, 1994; Heift, 1998; Loehr, 1998; James, 1998). For instance, translation students rarely include untranslated words. Instead, they either eliminate a portion of the passage or guess at the translation (James, 1998; Al-Onaizan, et al., 2000). Scorers were instructed, therefore, to treat untranslated words as lexical errors. In grading students, the scorers are instructed to not deduct multiple times for a recurring error. Since an MT system consistently makes the same errors, particularly lexical ones, the rules governing multiple errors was preserved and duplicate lexical errors were not counted twice in the same text.

The texts were then given to the scorers in order of text difficulty. Systems were presented in a uniform order of S1, S2, S3. The graders were given both the original source and the expert translation, although three of the four raters were fluent in French. The number of marks at each level was then counted and tabulated according to the weighting (Table 2). These marks were subtracted from the original word count and the result divided by the original word count for a final score. Because of the weighting scheme, it is possible for a system score to be negative.

4 Results

Given the type of data described, we expected this experiment to show three things. First, the scoring technique can be used to rank the systems successfully. Second, the text type does not affect the system ranking, but does indicate a relative MT capability. Third, the experiment indicates differences between scoring human language learners and scoring machines which are relevant to future adaptations.

4.1 Inter-scorer Reliability

Before looking at system rankings, the score variation, particularly for system S1, leads to questioning the scorers' reliability. Four raters were used to account for the fact that a particular scorer could be more lenient in general or on certain types of errors. The instructions to the raters are intended to minimize the subjective nature of the scoring by reducing it to a tagging task. Two items were used to explore rater effects on scores: post-interviews with the raters and inter-scorer reliability analyses using statistical techniques.

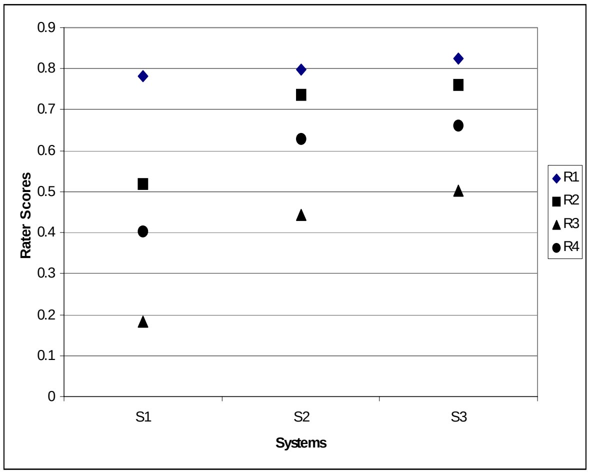

Figure 3: System Scores for Each Rater, Averaged on Text Type

Averaging across text type shows some consistency among the raters (Figure 3), though the magnitude of their differences varies, particularly for rater R1. In the post-interview, rater R1 said that she was an "easy grader" for foreign-language students and was the same way with translation systems. She did not follow the instructions as strictly as the other graders, since she stopped scoring once a threshold had been reached. Her consistently higher scores reflect this. Raters 3 and 4, on the other hand, are much closer to each other.

Inter-scorer reliability was measured by applying a Pearson correlation[5] to all rating points. The figure shows a consistency in ranking across raters. It can be seen, however, that rater R1 correlated weakly with the other scorers, with correlation ranging from 0.586 to 0.793. The inter-scorer correlation for the other scorers was high, ranging from 0.839 to 0.926. Because of the weak correlations for rater R1 and because she did not strictly follow guidelines, rater R1 results were eliminated from further calculations.

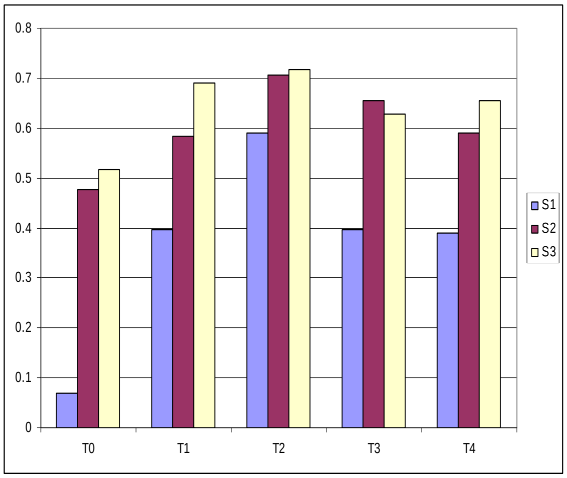

Figure 4: System Scores for Text Types

4.2 System Rankings

Aggregate scores are computed by a simple average across rater and text category. The aggregated scores indicate that the systems can be ranked from best to worst: S3, S2, S1 (Table 4; Figure 4). From these numbers and graphs, it can be seen that system S1 is clearly worse than S2 and S3, which is expected since system S1 is a lexical transfer system. Systems S2 and S3 are close to each other which reflects the fact that both are primarily syntactic transfer systems. A look at the scores for systems S2, S3 from the 1994 DARPA MT evaluation (White & O'Connell, 1994) for the language pairs shows that S2, S3 scores (Table 5) match those in this evaluation with S3 outperforming S2 slightly in adequacy (the 1994 scores are on a 0 to 1 scale). System S1 did not participate in the 1994 DARPA evaluation.

Table 4: System Scores

TEXT	S1	S2	S3
0	0.068	0.477	0.517
1	0.400	0.583	0.690
2	0.591	0.705	0.718
3	0.396	0.655	0.629
4	0.390	0.590	0.654
AVG	0.368	0.602	0.642

Table 5: DARPA 1994 Scores for S2, S3

SYSTEM	*ADEQUACY*	*FLUENCY*
S2	0.710220754	0.377458218
S3	0.789198359	0.503347118

Thus far in the analysis, the answer to the first question: "Can the method be used to successfully rank the systems?" is yes. The method accurately ranks the three systems and this ranking corresponds with an independent ranking of two of the systems. The ratings for text type T3 will be analyzed more closely in the next section since at this level, systems S2 and S3 are reversed in their ranking.

Table 6: Correlation of System Ranks.

	RANK S1	**RANK S2**	**RANK S3**
RANK S1	1.000	.0.926	0.843
RANK S2	0.926	1.000	0.924
RANK S3	0.843	0.924	1.000

4.3 System Ranking and Text Type

Can the method successfully rank the systems independent of text type? Additionally, can the method assign a minimal level of expertise for the MT systems? The scoring method distinguished between the quality of the systems by ranking them consistently on all but one text type, T3 (Figure 5), where systems S2 and S3 are ranked differently. In looking at the relative system rankings for each data point (rater, text type), the rankings correlate very well, independently of text type (Table 7). On the other hand, a proficiency level for MT cannot be assigned. The systems do not attain a passing grade at any of the levels.

[5] SPSS™ was used for data analysis.

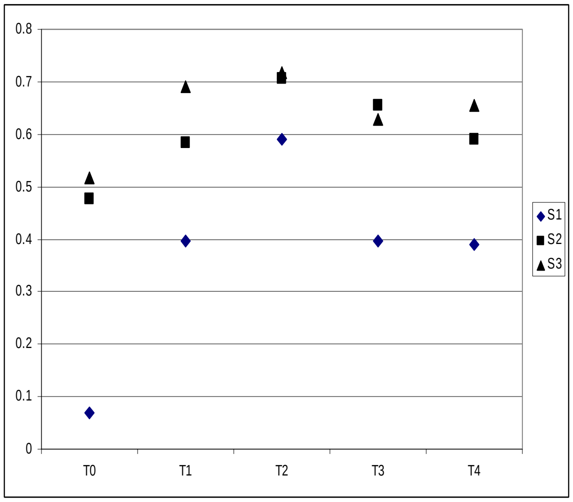

Figure 5: System Scores Averaged over Rater for Text Type

Unlike with students, the scores for MT are not decreasing as text difficulty increases. Given that the intent of the ILR scoring is to determine student level, the test assumes that a passing grade at one level implies that a student could pass at lower levels. Therefore, a student who successfully attains level two can be expected to pass levels one and zero without any difficulty. This assumption is not valid for scoring MT systems. It can be seen (Figure 6) that the system scores do not follow the profile described for human language learners. Instead, the scores reflect the fact that traditional MT systems are designed for news or technical text which best corresponds to level two texts. The pattern of scores cannot be accounted for by running order effects. The texts where shown from least complex to most complex and the expectation is that scores would decrease with each text since the scorers saw texts in order of difficulty. Instead, scores increase, peak and decrease. Therefore, running order is not a significant factor in the scores or system rankings.

Table 7: Level 0 Results for Each System, Rater

	S1	S2	S3
R2	0.224	0.552	0.584
R3	-0.06	0.328	0.344
R4	0.04	0.552	0.624
AVG	0.068	0.477	0.517

Table 8: Level 0 Breakdown of Errors per Category

	SYN	LEX	AWK	SLO	TOT
S1	1	150	44	3	198
S2	2	84	20	0	106
S3	1	77	23	0	101
TOT	4	311	87	3	405
% (TOT)	0.99	76.79	21.48	0.74	100

4.4 Level Zero Results

Expected results at this level are that given roughly equal lexicons, the systems would perform equally well, regardless of development methodology. Since this level has little syntactic or contextual information and no complete sentences (Lowe & Stansfield, 1988), even lexical transfer systems should perform on par with more complex systems. Because translation is into English, morphology errors would be minimal as well. The number of words in this text is 125, the lowest of all levels.

As can be seen from the total scores, none of these systems would have passed a level zero (T0) qualification exam (Table 7) since a passing grade would be greater than 0.90 and the best score is only 0.62. Due to the fact that this represents the least sophisticated language and should be easiest to translate, the result is surprising. S1 was expected to compete with S2 and S3 because syntactic constructions are minimal, with no complete sentences and rudimentary noun phrases. The syntactic systems perform significantly better than the lexical transfer system.

The most likely explanation is that the lexical transfer system has a smaller lexicon and/or phrasal dictionary. The proportion of closed class words in the output (16%) is smaller than for other levels (from 43% to 56%) and lends credence to this idea. The percentage of lexical category errors at 77% of 405 errors indicates this is the case. Additionally, an analysis of the translated texts and marked errors shows that the overall number of grammatical errors is small (Table 8). For the lexical transfer system, 38 of its 133 output words are marked as untranslated as opposed to less than five in each of the other outputs, another indicator that the S1 lexicon lacks coverage[6]. The error breakdown predicts that this level would be best suited for testing the breadth of the MT lexicon

[6] This count reflects not translated words and does not include incorrectly translated words.

due to the focus on word and phrase translation rather than grammatical transfer.

Table 9: Average Level 1 System Scores

	S1	S2	S3
R2	0.549	0.762	0.796
R3	0.128	0.335	0.524
R4	0.512	0.652	0.750
AVG	0.396	0.583	0.690

4.5 Level One Results

As noted in the previous section, if the scoring pattern for machines is the same as for scoring humans, the results expected at this level are that systems would perform slightly worse than at level zero. Additionally, because the texts have more sophisticated grammar to include complete sentences, the scores of system S1 should be lower than S2 and S3. Grammatical transfer systems would differentiate themselves from lexical transfer systems as the grammar penalty is twice the lexical entry penalty. A mitigating factor may be the percentage of common words and the increased number of words in the text. For level T1, the percentage of common words increases to 56% with the word count at 164. This means that systems could do as well as level zero here, although the introduction of more grammatical constructions may offset the gains from reduced lexical errors.

Table 10: Level 1 Category Breakdown of Scores

	SYN	LEX	AWK	SLO	TOT
S1	7	113	19	48	187
S2	6	82	10	14	112
S3	4	60	16	1	81
TOT	17	255	45	63	380
% (TOT)	4.47	67.11	11.84	16.58	100

Surprisingly, overall system scores increase rather than decrease (Table 9). While not attaining a passing grade, system scores are higher than at level T0. The number of errors has decreased from 405 to 380 despite an increase in word count from 125 to 164. The error breakdown shows the most likely cause of the improvements is related to the increased number of closed class words. The percentage of lexical errors has decreased (Table 10) from 77% to 67% since the percentage of common words has increased to 56%. The number of lexical errors has also decreased sufficiently to offset any increase in syntactic errors. While the lexical transfer system (S1) is worse than the syntactic systems (S2 and S3), it has narrowed the gap rather than the reverse. This can be explained by two factors. First, the lexical system does have a small phrasal lexicon. Second, since the languages involved are French and English, there is enough similarity in syntactic constructions to mitigate the syntactic penalties for lexical transfer.

4.6 Level Two Results

Results could be expected to show a slight decrease from level one, particularly for lexical transfer systems since the sentences in the text are longer with greater grammatical complexity than those in the earlier level. The level two texts are frequently factual news reports and contain full sentences. The number of sentences in level two has decreased to five as opposed to 32 in level one while the number of words is roughly the same at 155. Additionally, the texts represent a coherent set of paragraphs rather than unrelated vignettes. With language students, a score decrease would certainly be predicted. For MT systems, several factors come into play. With the increased syntactic difficulty, the syntactic errors should increase and systems S2 and S3 should perform much better than system S1. Second, with the percentage of closed class words at 46%, the lexical errors should remain low as in level one. Third, since MT systems are often designed for news texts (DARPA, 2003), they may be optimized for the texts at level two. Overall, a decrease in scores is predicted and a widening of the gap between lexical and syntactic systems is predicted.

Table 11: Level 2 Scores for Systems by Rater

	S1	S2	S3
R2	0.684	0.852	0.845
R3	0.452	0.594	0.671
R4	0.639	0.671	0.639
AVG	0.591	0.705	0.718

Table 12: Level 2 Category Breakdown of Scores

	SYN	LEX	AWK	SLO	TOT
S1	1	67	34	36	138
S2	0	57	23	0	80
S3	0	58	14	2	74
TOT	1	182	71	38	292
% (TOT)	0.34	62.33	24.32	13.01	100

Overall system scores continue to increase rather than decrease (Table 11). While not attaining a passing grade, system scores are highest at this level. S3 has the highest score, edging out system S2. Both systems outperform S1, although S1 also has its highest score at this level. The error count has decreased from 380 to 292 with roughly the same number of words. The error breakdown and word counts show one reason for better scores is related to the increased number of closed class words and repeat words. The ratio of closed class or repeated words is slightly greater than 50% of the total word count. The percentage of lexical errors has decreased (Table 12) from 67% to 62%. At the same time the count and percentage of syntactic errors has decreased for all systems, including the lexical transfer system. Two explanations have been described in the previous section, having to do with phrasal lexicons and syntactic similarities. Another more likely cause for this is an error attribution effect. The difficulty in attributing an error to an underlying cause is not unique in either language learning or MT evaluation (Schwind, 1995; Holland & Kaplan, 1995; Heift, 1998; Knight, 2000; Halliday & Briss, 1971; van Slype, 1979; Falkedal, 1994; Balkan et al., 1994; Taylor & White, 1998). Looking at the scoring of the judges, a tendency to attribute errors to lexicon difficulty appears. An examination indicates that errors were marked as lexical errors in multi-error situations.

4.7 Level Three Results

Expected results at level three are that systems should show a decrease from level two results. Texts here are not only factual reports, but also include speculation about future events and reasons behind the facts. The grammar is therefore more complex involving future tenses and modal verbs. The level of closed class words remains consistent at 49% while the word count decreases to 128. Like students, systems should score lower at this level due to the increased grammatical complexity. Additionally, the text includes more sophisticated named entities, referring to titles and organizations rather than people or places, requiring a more extensive lexicon.

Table 13: Level 3 Scores for Systems per Rater

	S1	S2	S3
R2	0.559	0.785	0.820
R3	0.223	0.539	0.496
R4	0.406	0.640	0.570
AVG	0.399	0.655	0.629

Table 14: Level 3 Category Breakdown of Scores

	SYN	LEX	AWK	SLO	TOT
S1	6	85	30	16	137
S2	1	55	15	7	78
S3	5	53	15	3	76
TOT	12	193	60	26	291
% (TOT)	4.12	66.32	20.62	8.93	100

The scores begin the decrease predicted by the increased complexity of the texts (Table 13). The number of grammatical errors has increased to twelve, despite the fact that the text length has decreased by 20% (Table 14). The total number of errors is the same as the previous level for a shorter text, showing an increased error rate. System S1 shows a marked decrease from level two, although the system scores more highly here than at level zero. This can be explained in part by the fact that the not-translated words in this level at eleven are higher than level two, but not as great in level zero at 38. Syntactic errors have arisen also due to more complex syntactic constructions.

Systems S2 and S3 do not follow all of the predicted trends. While their scores decrease from level two as predicted, their order is reversed. That is, system S2 scored more highly than S3 unlike at other levels. The main differentiation is due to the number of syntactic errors attributed to system S3 which is five as opposed to one for S2. A look at the texts reveals two possible reasons for this. First, the grammatical errors in system S2 could have been attributed to lexical errors in the phrase. Second, while both systems have advanced grammatical transfer, they could have different grammatical coverage. Due to the low number of words, even one or two different features covered per rater contributes to the difference between scores. After examining the scores, the most likely explanation is the first one. System S2 did have lexical errors marked where the S3 system had grammar errors marked.

4.8 Level Four Results

Since level four represents fluency in a language (Lowe & Stansfield, 1988), a sharp drop-off in the performance of all systems is expected. All of these systems have minimal semantic processing and therefore should have difficulty with the complex constructions and subtle expressions characteristic of this level. Partial sentences return at this level as needed for effect. In addition, abstract concepts such as **liberty** and **absurdity** make their appearance. The direct transfer systems should perform at their worst here as they lack even basic grammatical transfer and have no semantic resolution. On the other hand, the scoring guidelines do not include semantic errors as a separate category. Scorers generally marked these as lexical errors, therefore, lexical errors will increase. Given the merging of semantic errors with not translated words and other lexical choice errors, the direct transfer system may not suffer as badly as expected since the penalty for mistranslation and no translation are the same. The texts here have the highest number of words, 218, while the percentage of closed class words remains consistent at 56%. The number of named entities is relatively stable, although containing more personal and country names than the previous level.

Table 15: Level 4 Scores for Systems per Rater

	S1	**S2**	**S3**
R2	0.580	0.725	0.761
R3	0.177	0.422	0.482
R4	0.413	0.624	0.720
AVG	0.390	0.590	0.654

Table 16: Level 4 Category Breakdown of Scores

	SYN	**LEX**	**AWK**	**SLO**	**TOT**
S1	13	144	43	32	232
S2	11	96	30	4	141
S3	6	92	14	8	120
TOT	30	332	87	44	493
% (TOT)	6.09	67.34	17.65	8.92	100

At this level, the scores reflect expectations since they are lower than the scores from level three (Table 15). The systems are consistently ranked at this level, with the lexical transfer system much lower than the other systems. The scores for level four are not as bad as those for level zero, a function of the increased number of closed-class words. The decrease, therefore, reflects increased text difficulty as syntactic errors, 6.1% of the errors, play a larger part in the scores (Table 16). Although, as has been seen with level zero, the lexical errors are a significant proportion of errors at 68%. They are still the largest determiner in the final score.

Table 17: Error Breakdown by Text Type

	SYN	LEX	AWK	SLO	TOT	ERR / WORD
T0	4	311	87	3	405	0.360
T1	17	255	45	63	380	0.257
T2	1	182	71	38	292	0.209
T3	12	193	60	26	291	0.253
T4	30	332	87	44	493	0.251
TOT	64	1273	350	174	1861	0.266
%	3.44	68.40	18.81	9.35	100	

4.9 Error Breakdown

Not surprisingly, this experiment emphasizes that human learners and MT systems are good at different things linguistically. Lexical errors, which were nearly 70% of all errors (Table 17), were primarily untranslated words, something that learners would minimize. Given the fact that no semantic errors are called for explicitly in the scoring guidelines, it may be hypothesized that the lexical error category tends to encompass these errors as well. The hypothesis would lead to a call for more fine-grained criteria or additional categories into which semantic errors would fall.

Looking at the different levels of text, the errors per word shows the optimal level of texts for general MT scoring. By taking the total number of errors and dividing by the word count (per rater, per system for each text), the errors per word is calculated. This shows that the systems are optimized for texts characteristic of level two since the number of errors per word at 0.209 is less than the next closest level at 0.251. Finally the error rate in level T0 shows that general MT systems will have difficulty with the lower texts which are characterized by simplistic grammar and a low percentage of closed class words.

5 Lessons Learned

Unlike the later work of Clifford et al (2004) this effort treated MT as the student rather than the facilitator for students. While yielding informative results, this experiment falls short of desired evaluation properties in a number of key areas. As an evaluation, it is a human-intensive measure requiring selection of materials and multiple human scorers. Four scorers were the minimum necessary, even in the face of objective scoring guidelines. Second, while it does successfully rank the systems and does provide some feature information, it is not fine-grained enough due to the problem of error attribution. Third, a better test would be tailored to the kinds of mistakes that MT systems tend to make, indicating a slight change in the criteria to account for the error breakdown, such as mistranslated words versus untranslated ones. Finally, a wider range of test materials would be a better indicator of system ability.

References

Al-Onaizan, Y., Germann, U., Hermjakob, U., Knight, K., Koehn, P., Marcu, D. & Yamada, K. 2000. Translating with Scarce Resources. Proc. of AAAI '00.

Balkan, L., Arnold, D. & Meijer, S. 1994. Test Suites for Natural Language Processing. In Proceedings of Language Engineering Convention, CNIT, La Defense, Paris, France: 17-22

Carletta, J. C. 1996. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, 22(2), 249-254

Clifford, R., Granoien, N., Jones, D., Shen, W., Weinstein, C. 2004. The Effect of Text Difficulty on Machine Translation Performance – A Pilot Study with ILR-rated Texts in Spanish, Farsi, Arabic, Russian and Korean. In Proceedings of LREC 2004.

DARPA. 2003. Proceedings of the Workshop on MT Evaluation for Translingual Information Detection Extraction and Summarization (TIDES).

Falkedal, K. (ed.) 1994. Proceedings of the evaluators' forum, Les Rasses, ISSCO, University of Geneva, Geneva.

Halliday, T. & Briss, E. 1977. The Evaluation and Systems Analysis of the Systran Machine Translation System. Report RADC-TR-76-399, January, 1977. Rome Air Development Center, Griffiss Air Force base, New York. NTIS, ADA 036.070

Heift, G. 1998. Designed Intelligence: A Language Teacher Model. Unpublished Ph.D. thesis. Simon Fraser University.

Holland, V.M. & Kaplan, J. 1995. Natural Language Processing Techniques in Computer-Assisted Language Learning: Status and Instructional Issues. Instructional Science, 23: 351-380.

James, C. 1998. Errors in language learning and use. Exploring error analysis. London: Longman.

Knight, K. 2000. Machine Translation. Tutorial at ANLP/NAACL-2000

Loehr, D. 1998. Can Simultaneous Interpretation Help Machine Translation. In Proceedings of AMTA 1998.

Lowe, Jr., P. & Stansfield, C. 1988. Introduction. In P. Lowe, Jr. & C. Stansfield, eds., Second language Proficiency Assessment: Current Issues. Center for Applied Linguistics. Prentice-Hall Regents.

Michaud, L. & McCoy, K. 1999. Modeling User Language Proficiency in a Writing Tutor for Deaf Learners of English. In M. Olsen, ed., Computer-Mediated Language Assessment and Evaluation in Natural Language Processing, Proceedings of a Symposium by ACL/IALL. University of Maryland, p. 47-54

Schwind, C. 1994. Error Analysis and Explanation in Knowledge-Based Language Tutoring. In L. Appelo & F.M.G. de Jong eds., Computer-Assisted Language Learning, Proceedings of the Seventh Twente Workshop on Language Technology.

Schwind, C. 1995. Error analysis and Explanation in Knowledge-based Language Tutoring. Computer-Assisted Language Learning. 8(4). 295-324

Taylor, K. & White, J. 1998. Predicting what MT is Good for: User Judgments and Task Performance. Proceedings of AMTA-98.

Van Slype, G. 1979. Critical Methods for Evaluating the Quality of Machine Translation. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR-19142. Bureau Marcel van Dijk

Vanni, M. & Reeder, F. 2000. How Are You Doing? A Look at MT Evaluation. In Proceedings of AMTA-2000. October

White, J., & O'Connell, T. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proceedings of the 1994 AMTA