Formal v. Informal:

Register-Differentiated Arabic MT Evaluation in the PLATO Paradigm

Keith J. Miller* **Michelle Vanni**[§]

The MITRE Corporation*
McLean, VA 22102
keith@mitre.org

U.S. Army Research Laboratory[§]
Adelphi, MD 20785
mvanni@arl.army.mil

Abstract

Tasks performed on machine translation (MT) output are associated with input text types such as genre and topic. Predictive Linguistic Assessments of Translation Output, or PLATO, MT Evaluation (MTE) explores a predictive relationship between linguistic metrics and the information processing tasks reliably performable on output. PLATO assigns a linguistic signature, which cuts across the task-based and automated metric paradigms. Here we report on PLATO assessments of clarity, coherence, morphology, syntax, lexical robustness, name-rendering, and terminology in a comparison of Arabic MT engines in which *register* differentiates the input. With a team of 10 assessors employing eight linguistic tests, we analyzed the results of five systems' processing of 10 input texts from two distinct linguistic registers for a total of 800 data sets. The analysis pointed to specific areas, such as general lexical robustness, where system performance was comparable on both types of input. Divergent performance, however, was observed for clarity and name-rendering. These results suggest that, while systems may be considered reliable regardless of input register for the lexicon-dependent triage task, register may have an affect on the suitability of MT systems' output for relevance judgment and information extraction tasks, which rely on clearness and proper named-entity rendering. Further, we show that the evaluation metrics incorporated in PLATO differentiate between MT systems' performance on a text type for which they are presumably optimized and one on which they are not.

1. Introduction

The information-processing task performed on the output of a machine translation system is closely associated with the text type submitted as input to the system. Genre, complexity and topic are but a few of the characteristics with respect to which input can be differentiated. Media also play a part as ASR and OCR serve to degrade input regardless of other features.

The PLATO MT Evaluation (MTE) program explores the possibility of a predictive relationship between discrete, well-defined MTE metrics and the tasks that can be reliably performed with MT output. Scores on PLATO assessments constitute a signature to be correlated with these tasks and with automated metrics. PLATO assesses clarity, coherence, morphology, syntax, lexical robustness, named-entity rendering and adequacy (Miller & Vanni, 2002; Vanni & Miller, 2002).

In the present work, we use PLATO metrics in an operational comparison of Arabic MT engines in which the level of formal *register* was the input-differentiating feature. With a team of 10 assessors, we tested five Arabic-English MT systems, using 10 input texts and eight linguistic tests.

Analysis of the 800 sets of scores revealed specific areas, such as general lexical robustness, where system performance was comparable on both types of input. Divergent performance occurred on assessments of clarity, name rendering and domain terms.

Assuming a correlation between system performance on the Dictionary Update assessment and user performance on the triage task, results suggest that the Arabic MT systems evaluated may be considered reliable on both types of input for the task of triage. Suitability for the tasks of relevance judgment, information extraction and topic identification, however, which rely on clearness, proper named-entity rendering, and domain term accuracy, respectively, would be more reliably judged with consideration given to the register of the input text.

[§]The opinions, interpretations, recommendations and conclusions expressed herein are those of the author and are not necessarily endorsed by the USG.

2. The PLATO Research Program

In the Predictive Linguistic Assessments of Translation Output (PLATO) MT Evaluation (MTE) program, we explore the possibility of a predictive relationship between discrete, well-defined MTE metrics and the specific information processing tasks that can be reliably performed with MT output. PLATO has developed linguistic assessments, scores which constitute a signature to be correlated with specific tasks and with automated metrics.

2.1. Overview

PLATO consists of seven traditional measures of quality, informed by the International Standards for Language Engineering (ISLE) and its Framework for Evaluation of Machine Translation in ISLE (FEMTI) (Hovy, et al. 2003). It also includes, for this study, a DARPA-style Adequacy test (White and O'Connell 1994). The approach draws inspiration from both the task-based (Church & Hovy 1993; Taylor and White 1998; Jones, et al. 1994, 2004; Weinberg, et al. 2005) and automated (Papineni et al., 2001, 2002; Melamed 2003; Lavie 2004, Snover, et al. 2005) MTE paradigms.

Current focus is in three areas: relating patterns in assessment scores to performable tasks, measuring and optimizing inter-rater agreement in the performance of the metrics and, finally, automating the assessments. It is expected that this will be done through recognition of correlations with other, possibly existing, metrics. For this reason, the PLATO program work can also be viewed as preliminary design for a program that would report on MT usability by means of meta-information on the output.

2.2. Semantics-Based Assessments

The first and last assessments in the PLATO suite for this study are Clarity and Adequacy. Each focuses on sentence-level expression. We measure the lucidity of expression in the output with the Clarity test; we measure the extent to which the output meaning expresses the input meaning with the Adequacy test. The latter is modeled on the eponymous 1994 DARPA test.

The Clarity test is performed without reference to a source text or reference translation. It is a snap judgment of the degree to which some discernible meaning can be assigned to the sentence. The measure ranges between "0" for output the meaning of which is not apparent, even after some reflection, and "4" for output meaning which is perfectly clear on first reading.

By contrast, in the Adequacy test, assessors compare output meaning with the meaning expressed in a reference translation. The fidelity of the output is measured on a scale from "1" for output which only minimally matches reference meaning, to "5" for output that matches all reference meaning.

2.3. Structure-Based Assessments

Morphology, Syntax and Coherence assess the structure of the output at the word-, sentence- and discourse-levels. Coherence is the second assessment in the suite, followed by the Morphology and Syntax assessments. Coherence is measured after Clarity to prevent bias and attenuate possible training effects.

In the Morphology test, assessors are presented with an output sentence and asked to identify word formation errors. The score is calculated as one minus the ratio of corrections to the number of inflectable words.

The Syntax test asks assessors to transform output sentences into grammatical sentences by making a minimal number of corrections, reorderings, deletions and additions. The assessment is similar to the DARPA-sponsored edit-distance metric, TER, being used in the GALE research program (Snover, et al. 2005). The PLATO syntax score for each sentence, in contrast to the GALE metric, is simply calculated as one minus the ratio of the number of changes to the number of words in the sentence.

With Coherence testing, assessors consider output sentences one-at-a-time in context and determine if one of the functions defined in Mann and Thompson's (1981) Rhetorical Structure Theory (RST) can plausibly be assigned to it. Thus, sentence scores are binary with RST constraining the set of assignable functions.

2.4. Lexicon-Based Assessments

The last three tests in the PLATO suite, Dictionary Update, Name Rendering and Domain Terminology, measure output quality at the lexical level. All depend on consultation of a human reference translation to provide ground truth. Since these measures are fairly deterministic, maximal inter-rater agreement is the norm.

In the Dictionary Update test, assessors identify untranslated words in output. Name Rendering and Domain Terminology require an additional step of creating ground-truth lists of names and domain terms against which to compare the output text. To facilitate eventual automation, exact match is required for a judgment of correct to be made. For all three, scores are calculated as one minus the percentage of corrections.

2.5. Previous Work

PLATO assessments were given preliminary validation on the output of systems handling source languages which are both close in structure to the target as well as highly divergent (Vanni & Miller, 2002). In both cases, the assessments appeared to rank system output quality. More recently, with refinement of metrics and guidelines, PLATO Clarity assessments achieved a joint probability of inter-rater agreement, or reliability, in the .8 or "good" range (Miller & Vanni, 2005).

3. Register-Differentiated Input

As an evaluation mechanism, PLATO aims to predict where problems in output will occur. For the present study, we expected that Arabic MT system performance would vary with the type of language input to the system. Since most systems are trained on MSA, we hypothesized that output from MSA input would achieve generally higher assessment scores than that from input which is not standard. We also wanted to test our assumption that more fine-grained predictions would depend on the exact type of variations found in the non-standard input.

Pursuant to these aims, we used as input both standard (MSA) and non-standard Arabic text such as can be found in written electronic discourse. Moreover, phenomenon-specific input differences provided a focus for making MT performance predictions. That is, the differences helped pinpoint the assessments on which nonstandard output would score relatively well or relatively poorly.

3.1. Variation: Dialect and Genre

It is a widely held belief that the Arab-speaking world presents a diglossic environment in which one dialect is used for more public purposes and another is used in more familiar situations. However, stylistic variations consisting of combinations of shared and mixed dialect features, as well as new genre distinctions, actually conspire to create more of a multiglossic continuum (Freeman 2002). Moreover, widespread internet communication has further complicated the picture by capturing in written form linguistic variation hitherto principally lost in speech streams.

In fact, for language in general, the web has inevitably led to the emergence of new *language varieties* or "system[s] of linguistic expression whose use is governed by situational factors" (Crystal 2001:6). These may include factors relating to geography, as with dialects, or to style and form, as with genres.

Dialect can be defined as:

> A distinct form of a language (or other communication system) that differs from other forms of that language in specific features (pronunciation, vocabulary, and/or grammar), possibly associated with some regional, social, or ethnic group, but that is nevertheless mutually intelligible with them (Aikmajian, et al. 1988).

While most features that distinguish dialects can be readily ascertained in spoken genres, such as conversations and speeches, they are obscured in textual form by the Arabic script. The pronunciation of certain consonants and the placement and quality of vowels, along with indication of syllable boundaries, are both, in addition to specific lexical cues, good indicators of dialect and needed for well-founded dialect determination.

Keeping in mind that the term *genr*e refers to "a category of […] literary composition characterized by a particular style, form or content," (Merriam-Webster 2006) what we surmise from our observations serves as a set of operating premises.

The first is that dialects can be distinguished with a high degree of reliability among spoken genres. We also know that they are usually mixed with MSA in specific contexts, (e.g. broadcasts, such as talk shows).

Next, we understand that dialect use can be predicted with a high degree of reliability for traditional written genres. That is, for example, that Classical Arabic will be used in koranic texts, MSA will be used in newswire and, regional dialects will be used for locally distributed publications, such as flyers and brochures.

Finally, the web introduces a new medium that blurs previously crisp dialect and genre distinctions by conflating the evidence for them. In this new environment, deprived of the phonological evidence that characterizes traditionally spoken language varieties, one is left with infrequently occurring, and often shared, lexical and syntactic cues, which allow only for probabilistic dialect determinations.

Moreover, as is often noted, asynchronous web-based interaction burdens authors with negotiating opposing tensions, such as public v. familiar and polite v. personal, when making style and content choices. These concerns, previously guided by traditional genre definitions, are, in web forums, handled idiosyncratically. We have observed a range of common structural and content-based features, varying only in degree, in three different web contexts. These include discussion lists, weblogs and on-line editorials, and are detailed in Section 3.2.

3.2. Formal v. Informal Register

Given that characterizing Arabic web text input to MT tends to be an inexact science, we deemed "formal v. informal" to be a more apt distinction, hinging as it does on combined features of dialect and genre. The descriptive linguistics literature confirms this intuition as Lyons (1981) notes, "in so far as stylistic variation is determined, or conditioned, by the social context, it falls within the sociolinguistic concept of *register*."

Formal MSA is a common language variety, which is used in the traditional genres of scientific and news reporting, among others. For this study, *the formal register* input category consisted of MSA newswire text.[1]

The *informal register* input category consisted of samples taken from three web contexts. Each context represents a specific *variety*, in Crystal's terms, of Arabic language. Table 1 gives evidence of these varieties.

[1] In considering formal register, we also recognize the Classical Arabic of religious and literary texts, an unlikely candidate for MT except as quotations interspersed in other texts; hence, beyond the scope of our study

First in the informal category is a variety of MSA containing predictable, non-dialect-specific features of grammar and vocabulary, i.e., pan-Arab lexical items with turns of phrase and reduced phonological and case-based morphological affixation not found in formal MSA. A spoken standard, this variety is widely studied and taught as Formal Spoken Arabic (Ryding & Mehall 2005).

On-Line Editorial	**Discussion List**	**Internet Web log**
Features common to many dialects	*Lexical features of specific dialect*	*Morph features of specific dialects*
illi (v. alladhi) [Rel. Pronoun]	yishuuf 'to see'	qaa'id [Pres. Continuous]
bi- [pre verbal] [Indicative Mrkr]	'alashaan 'on account of'	madrayt 'still yet'
sa-tafriD 'she will force'	sa- [Future Marker]	ma- [Negation]

Table 1. Informal Register Input Category: Evidence for Constitutive Language Varieties

Next in the informal category are two dialect-based varieties. *Discussion lists* contain obvious lexical signs of dialect, to include the forms of 'to see' and 'on account of' as well as other cues, such as morphological and phonological reductions. *Web logs* were found to contain many examples of dialect-specific morphological forms marking verb tenses and other features of aspect and negation. Table 1 shows some examples of this evidence. We refer to the two non-MSA informal varieties as Dialect Blog because they feature characteristics of specific dialects and are structurally less well-defined than the web-editorial genre which we refer to as MSA Blog.

Note that common to both informal varieties are features on which it is unlikely that MT systems have been trained, especially the non-standard morphological affixation and the forms containing phonological reductions. Thus, we would predict discrepancies between PLATO morphology assessment scores on output from formal input and those on output from informal input.

4. Experimental Design

4.1. Systems

We used several commercially available Arabic-to-English MT systems. Among these were statistical and hybrid engines. Some systems were run at multiple settings, thus simulating five different systems.

4.2. Blocking

The input consisted of five samples from each of the two input categories of Formal and Informal. The five samples from the Formal category were MSA newswire and the five samples from the Informal category consisted of two MSA Blog articles and three Dialect Blog articles. Thus 10 texts were fed to five systems to produce fifty outputs, each representing a unique input + system combination. Table 2 shows designations for each output with outputs from the Informal input category appearing in shaded cells.

System	S1	S2	S3	S4	S5
Input					
SN1	SN1s1	SN1s2	SN1s3	SN1s4	SN1s5
SN2	SN2s1	SN2s2	SN2s3	SN2s4	SN2s5
SN3	SN3s1	SN3s2	SN3s3	SN3s4	SN3s5
SN4	SN4s1	SN4s2	SN4s3	SN4s4	SN4s5
SN5	SN5s1	SN5s2	SN5s3	SN5s4	SN5s5
SB1	SB1s1	SB1s2	SB1s3	SB1s4	SB1s5
SB2	SB2s1	SB2s2	SB2s3	SB2s4	SB2s5
DB1	DB1s1	DB1s2	DB1s3	DB1s4	DB1s5
DB2	DB2s1	DB2s2	DB2s3	DB2s4	DB2s5
DB3	DB3s1	DB3s2	DB3s3	DB3s4	DB3s5

Table 2. Output designations. First letter indicates dialect: MSA (S) or Dialect (D); second is genre: news (N) or Blog (B); third digraph designates system (s1-5).

After several iterations of training and guideline refinement, ten assessors, with expertise in either linguistics, copy editing, or language pedagogy, viewed 10 outputs each. Systems were assigned a unique output-assessor pairing for each category of input, creating a blocking which allowed two assessors to view each output. Table 3 shows output-assessor ordering for each system within the two blocking sequences, **F**(ormal)**X**--**I**(nformal)**X** and **I**(nformal)**X**--**F**(ormal)**X**, with X indicating system number. In this way, no assessor saw more than one output from same input and each assessor evaluated output from a unique set of input-plus-system combinations. Moreover, viewing order was randomized for each assessor.

Suites of eight PLATO assessments were performed on each assigned output, producing 800 individual sets of scores. For a given system, the two assessors' scores on the output, produced from each of the five input category texts, were averaged and results recorded for further analysis.

5. Results

5.1. Comparing Registers

Table 4 shows average scores for PLATO assessments performed on output from five Arabic-English MT systems fed Formal (F) and Informal (I) register category data. The overall superior performance of the systems on the Formal data is readily apparent. The highest average differences were on the Coherence (60.7%), Clarity (38.75%), and Domain Terminology (27.84%) assessments. Lowest was Dictionary Update (6.46%).

5.2. Morphology and Proper Names

Notable among our results were scores on the Morphology and Proper Name Assessments. Predictably, the average Morphology score over all systems on Formal input (.89) exceeded (by close to 15%) the same score calculated on output from Informal input (.75). Moreover, differences in Morphology scores for the poorer systems on Informal input approached .20 with the standard deviation among scores on this output more than doubling that of the Formal category. These marked differences suggest that PLATO accurately reflects the inability of standard MT systems to account for the widely varying morphological phenomena common to Informal Arabic.

The Proper Name Assessment reflected a similar predicted baseline of performance. However, scores were generally much lower. The average score for systems on output resulting from Formal input was .50, with the average for Informal input being .29. Differences in scores for more poorly performing systems reached the 30% level. Standard deviations (StdDev) among scores were also concomitantly greater for the output from Informal input on the Proper Name Assessment, with a StdDev of 4.8 among the scores on output from Formal input and 13.4 for the scores on output from Informal input. Thus, treatment of Arabic names showed itself to be an area where Arabic-English system developers could focus to great advantage.

Block Sequence	F1-I1	F2-I2	F3-I3	F4-I4	F5-I5	I1-F1	I2-F2	I3-F3	I4-F4	I5-F5
Assessor										
A01	SN1s1	SN2s2	SN3s3	SN4s4	SN5s5	SB1s1	SB2s2	DB1s3	DB2s4	DB3s5
A02	SN2s1	SN3s2	SN4s3	SN5s4	SN1s5	SB2s1	DB1s2	DB2s3	DB3s4	SB1s5
A03	SN3s1	SN4s2	SN5s3	SN1s4	SN2s5	DB1s1	DB2s2	DB3s3	SB1s4	SB2s5
A04	SN4s1	SN5s2	SN1s3	SN2s4	SN3s5	DB2s1	DB3s2	SB1s3	SB2s4	DB1s5
A05	SN5s1	SN1s2	SN2s3	SN3s4	SN4s5	DB3s1	SB1s2	SB2s3	DB1s4	DB2s5
A06	SB1s1	SB2s2	DB1s3	DB2s4	DB3s5	SN1s1	SN2s2	SN3s3	SN4s4	SN5s5
A07	SB2s1	DB1s2	DB2s3	DB3s4	SB1s5	SN2s1	SN3s2	SN4s3	SN5s4	SN1s5
A08	DB1s1	DB2s2	DB3s3	SB1s4	SB2s5	SN3s1	SN4s2	SN5s3	SN1s4	SN2s5
A09	DB2s1	DB3s2	SB1s3	SB2s4	DB1s5	SN4s1	SN5s2	SN1s3	SN2s4	SN3s5
A10	DB3s1	SB1s2	SB2s3	DB1s4	DB2s5	SN5s1	SN1s2	SN2s3	SN3s4	SN4s5

Table 3. Blocking of output texts showing that each output was viewed by two assessors and that no output from the same input was viewed twice by any one assessor.

Metric	Clarity		Coherence		Syntax		Morph		Dictionary Update		Domain Terms		Proper Names	
Input Type	F	I	F	I	F	I	F	I	F	I	F	I	F	I
System														
S1	72.75	43.25	94.60	45.50	92.30	84.50	92.20	82.70	98.60	92.30	48.10	20.80	44.40	6.70
S2	64.75	17.75	96.30	25.00	79.50	70.00	82.90	65.20	98.40	92.40	42.70	22.50	56.80	39.20
S3	71.00	29.25	96.20	34.90	88.70	73.70	88.20	69.00	98.70	91.50	47.10	15.20	53.00	38.30
S4	69.75	29.75	97.40	39.50	89.30	79.70	90.00	78.00	99.30	92.30	46.40	17.40	47.30	34.20
S5	68.25	32.75	95.90	32.00	91.90	78.90	91.20	82.50	99.10	93.30	46.70	15.90	49.60	27.50

Table 4. Average PLATO results on seven assessments, given inputs from Formal (F) and Informal (I) register categories, on output from five Arabic-to-English MT systems.

As noted above, PLATO creates unique linguistic signatures for MT systems. Figure 1 compares the signatures for the five systems on Informal input and shows performance which is roughly equivalent across systems for Domain Terms and consistently superb for Dictionary robustness. Signature components for Clarity, Coherence, Syntax, Morphology, and Proper Names varied widely, however. Matching the linguistic requirements of tasks to such specific strengths and weaknesses of systems in handling input types of interest is basic to providing output tagged for task suitability.

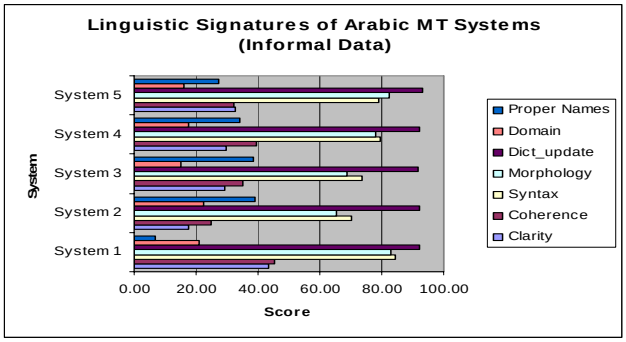

Figure 1. System-level PLATO Linguistic Signatures for Informal Data

5.3. "System One"

Results of the PLATO comparison of quality of MT output from two different input registers clearly indicate that currently available Arabic-English systems are generally producing outputs from Formal input of a quality which greatly exceeds that produced from Informal input. This difference is more pronounced than might be expected and, in the PLATO program, can be localized. For example, System One results in Figure 2 show that its general lexicon is adequately robust for the handling of both input types but that the output produced from Formal input was much clearer and more coherent than that produced from Informal input.

6. Related Work

MTE research into automated methods for quantifying the linguistic qualities of output has resulted in the METEOR (Lavie et al. 2004) and the TER (Snover et al., 2005) systems, mentioned in Section 2. Weinberg (2005) and Tate, et al. (2005) both sought to predict task performance on MT output using existing automated metrics. Finally Henderson (2004) is pursuing machine learning approaches for automatically determining the interpretability of an output text in the TIRS program.

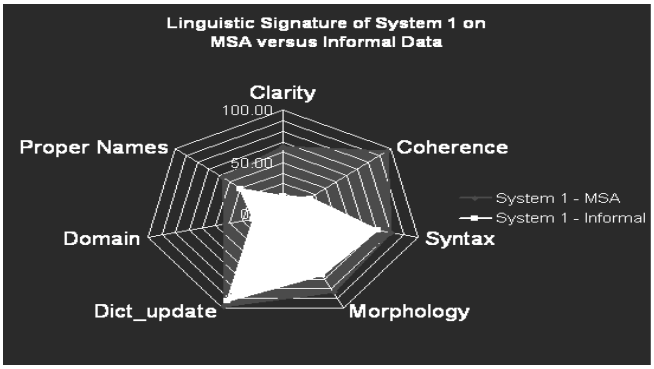

Figure 2. System One PLATO Signatures for Formal v. Informal Input Texts

7. Conclusions and Future Work

In this study, PLATO metrics correctly predict that, based on the widely divergent morphological features of the two input categories, the quality of the MT output from each would reflect similar differences. Our hypothesis, that PLATO assessments are sensitive enough to reflect differences in MT systems' performance on inputs as linguistically varied as Formal Arabic newswire and Informal web data, has been reinforced. The linguistic signatures produced by the same system, when operating on these two data types, are markedly distinct.

This result demonstrates that PLATO's metrics can pinpoint differences in output. The nature of these differences enables evaluation to span the user-developer divide, acting as a simultaneous service, both as an advisory for users and an indicator of areas of improvement for developers.

Based on these output characterizations, it is plausible that PLATO's insistence on standard linguistic criteria in assessment may also cut across differences among strategies for evaluating individual components, when employed, in possible future work, for evaluating MT in embedded contexts.

Degradation from preprocessing, such as that occurring in ASR and OCR output, will further alter the linguistic nature of the output from that which would have been produced from clean input text. PLATO signatures are equipped to capture those differences so as to, among other things, indicate the appropriateness of the output for downstream processing by, for example, information extraction, retrieval and summarization processes.

8. Acknowledgements

We would like to thank Dr. Andrew T. Freeman and Dr. Dan Parvaz for their expert help with analysis of the Arabic web data.

9. References

Aikmajian, A., R. A. Demers and R. M. Harnish. (1988). Linguistics: An Introduction to Language and Communication, Second Edition. Cambridge, MA: MIT Press. 521.

Church, K. and E. Hovy. (1993). Good Applications for Crummy Machine Translation. *Machine Translation* 8: 239-258.

Crystal, D. (2001). *Language and the Internet*. Cambridge: CUP.

Freeman, A. (2002). *In Search of a koine in Sana'a*. Unpublished Ph.D. Dissertation. Departments of Linguistics and Near Eastern Studies. University of Michigan.

Henderson, J. (2004). The Text Interpretability Rating Scale. Presentation at Panel on Trends in MT Evaluation. Conference of the Association for Machine Translation in the Americas (AMTA) Washington, DC.

Hovy, E., M. King and A. Popescu-Belis. (2003). Principles of Context-Based Machine Translation Evaluation. *Machine Translation* 17:1. 43–75.

International Standards for Language Engineering. (2000). ISLE Classification MT Evaluations. http://www.isi.edu/natural-language/mteval. In *Proceedings of Hands-on Workshop on Machine Translation Evaluation. Association for Machine Translation in the Americas*, Cuernavaca, Mexico.

Jones, D. and G. Rusk. 2000. Toward a Scoring Function for Quality-Driven Machine Translation. In *Proceedings of COLING-2000*.

Lavie, L., K. Sagae and S. Jayaraman. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In Machine Translation: From Real Users to Research, *Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, Washington, D.C.

Lyons, J. (1981). *Language and Linguistics: An Introduction.* Cambridge:CUP. 292.

Mann, W.C. and S.A. Thompson. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. Text 8:3. pp. 243-281.

Melamed, I, R. Green, and J. Turian (2003) *Precision and Recall of Machine Translation*. In Proceedings on NAACL/HLT, Edmonton, Canada.

Miller, K.J. and M. Vanni (2002). Scaling the ISLE Taxonomy: Development of Metrics for Multi-Dimensional Characterization of Machine Translation Quality. In *Proceedings of MT Summit VIII*. Santiago de Campostela, Spain.

Miller, K. J. and M. Vanni. (2005). Inter-rater Agreement Measures and the Refinement of Metrics in the PLATO MT Evaluation Paradigm. In *Proceedings of MT Summit X*, Phuket, Thailand.

Papineni, K., S. Roukos, T. Ward and W. Zhu. (2001). BLEU: A Method for Automatic Evaluation of Machine Translation, IBM Research Report, RC22176, September 2001.

Papineni, K., S. Roukos, T. Ward, J. Henderson, and F. Reeder. (2002). Corpus-Based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results. *Proceedings of the Human Language Technology Conference*.

Ryding, K.C. and D.J. Mehall. (2005). *Formal Spoken Arabic*. Washington, DC: Georgetown University Press.

Snover, M., B.J. Dorr, R. Schwartz, J Makhoul, L. Micciula and R. Weischedel. (2005). A Study of Translation Error Rate with Targeted Human Annotation. Technical Report LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, College Park.

Tate, C., C. Voss, B. J. Dorr and E. Slud. (2005). Toward a Predictive Statistical Model of Task-based Performance Using Automatic MT Evaluation Metrics. In *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Ann Arbor, MI.

Taylor, K. and J. White. (1998). Predicting What MT is Good for: User Judgments and Task Performance. In: *Proceedings of the 1998 conference of the Association of Machine Translation in the Americas*. 364-373.

Vanni, M. (1998). Evaluating MT Systems: Testing the Feasibility of a Task-Diagnostic. *Proceedings of the Association for Information Management (ASLIB): Translating and the Computer 20*.

Vanni, M. & Miller, K.J. (2002). Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Metrics across Languages. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*. Las Palmas de Gran Canaria, Spain. 1254-1262.

"Genre." Merriam-Webster On-Line Dictionary. (2006). http://www.m-w.com/dictionary/genre. (27 Feb 2006).

Weinberg, A. (2005). Machine Translation Evaluation at the Center for Advanced Study of Language. *Workshop on Intelligence Analysis*, Mitre Corporation.

White, J. and O'Connell T.A. (1996). The ARPA MT evaluation methodologies: evolution, lessons and future approaches. *Proceedings of the 1994 Conference, of the Association for Machine Translation in the Americas.*