Evaluation principles and objectives: ISLE, EAGLES (A play in three acts)

ELRA / ELDA HLT Evaluation Workshop

December 1-2, 2005
Malta

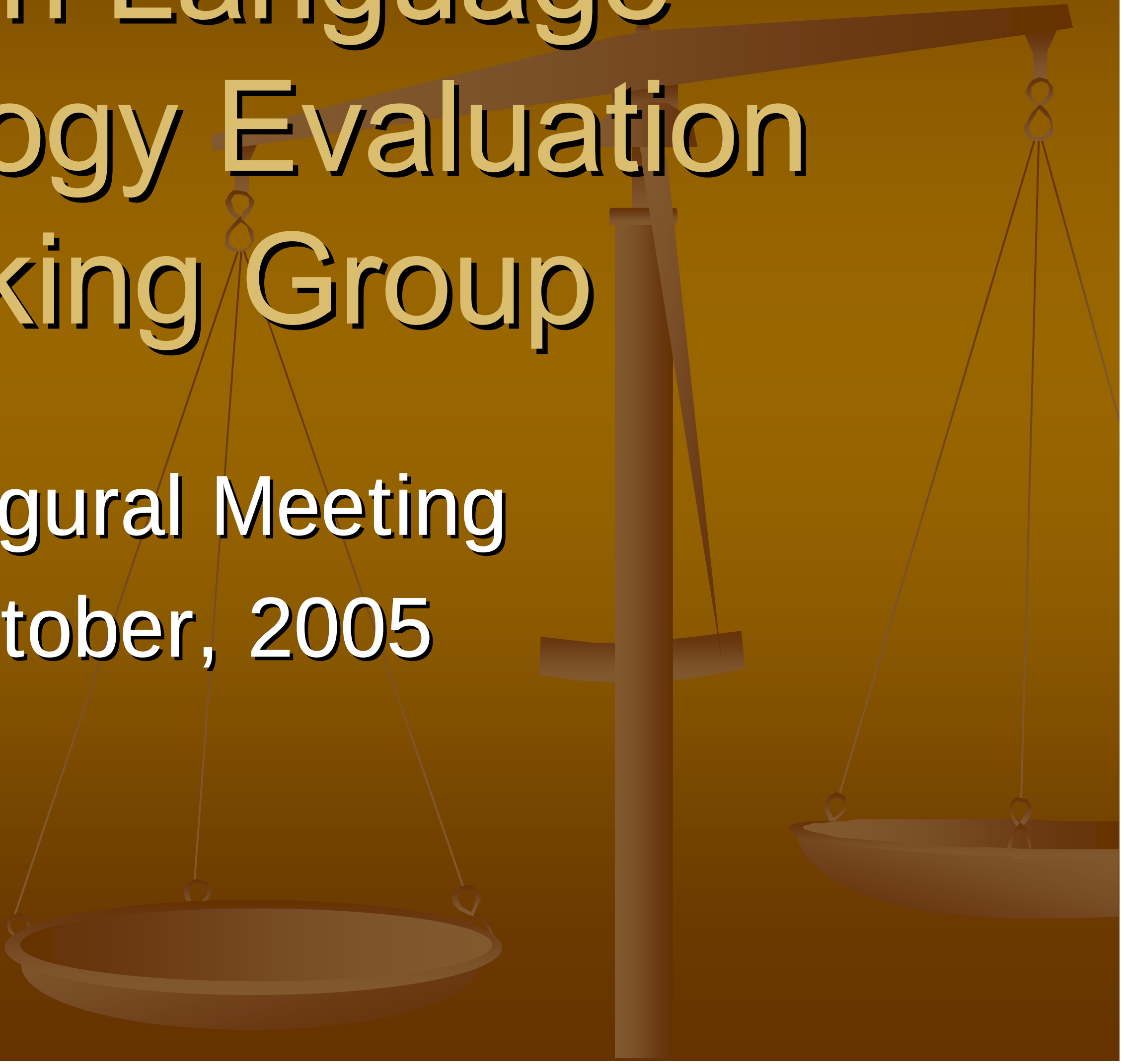
Keith J. Miller
The MITRE Corporation

Act 1: Exposition

Maghi King: « A question for this workshop: »

- How can we build on what we have learnt in order to
  - deploy effectively knowledge and experience gained
  - share experience and insights as they develop
  - build bridges to other evaluation communities
  - meet new challenges

Human Language Technology Evaluation Working Group



Inaugural Meeting

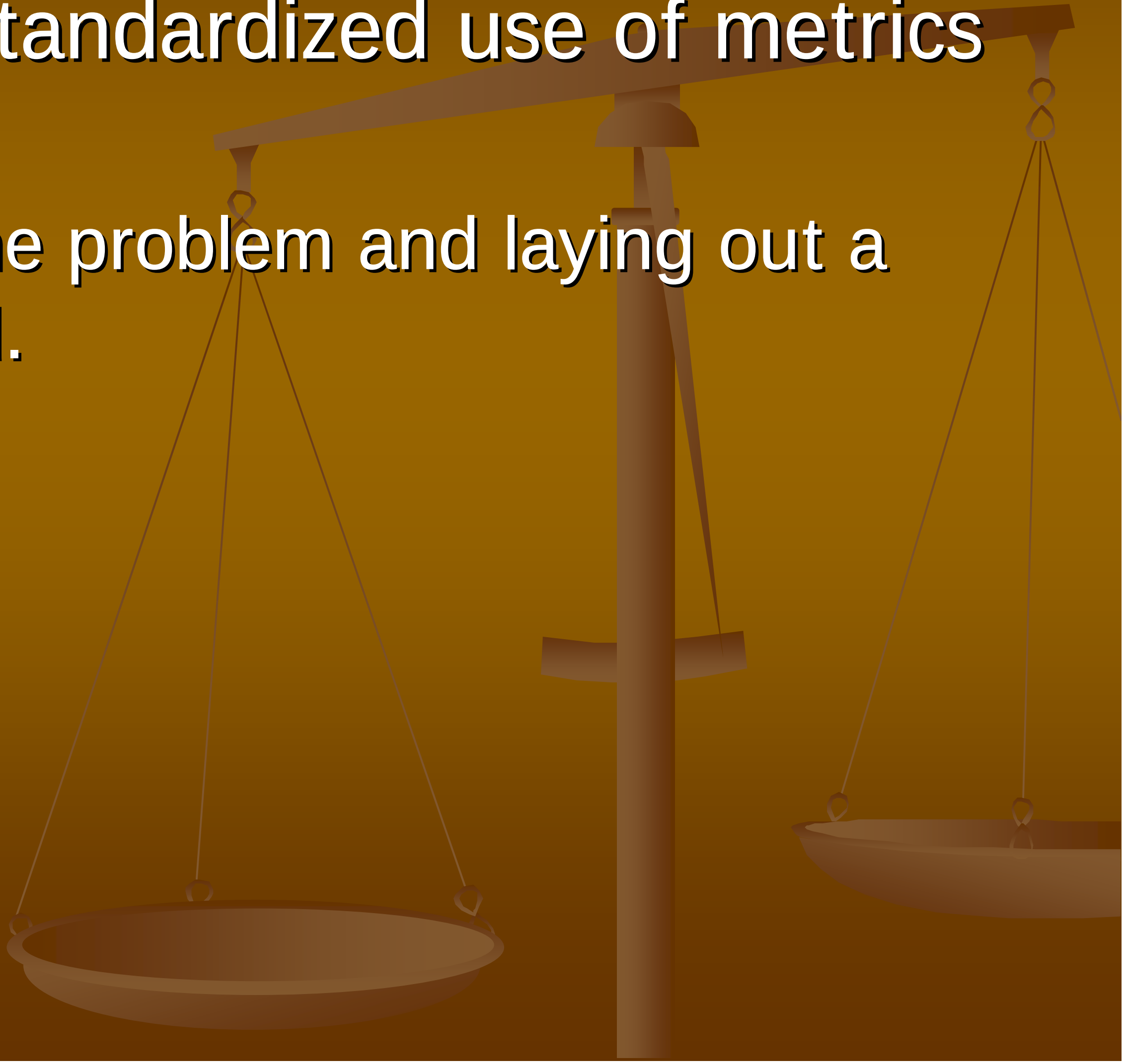
October, 2005

Agenda

- Meeting Purpose
- What Makes a Good Evaluation?
- An Evaluation Framework
- Overview of NLP Technology Metrics,
  …
- Next Steps

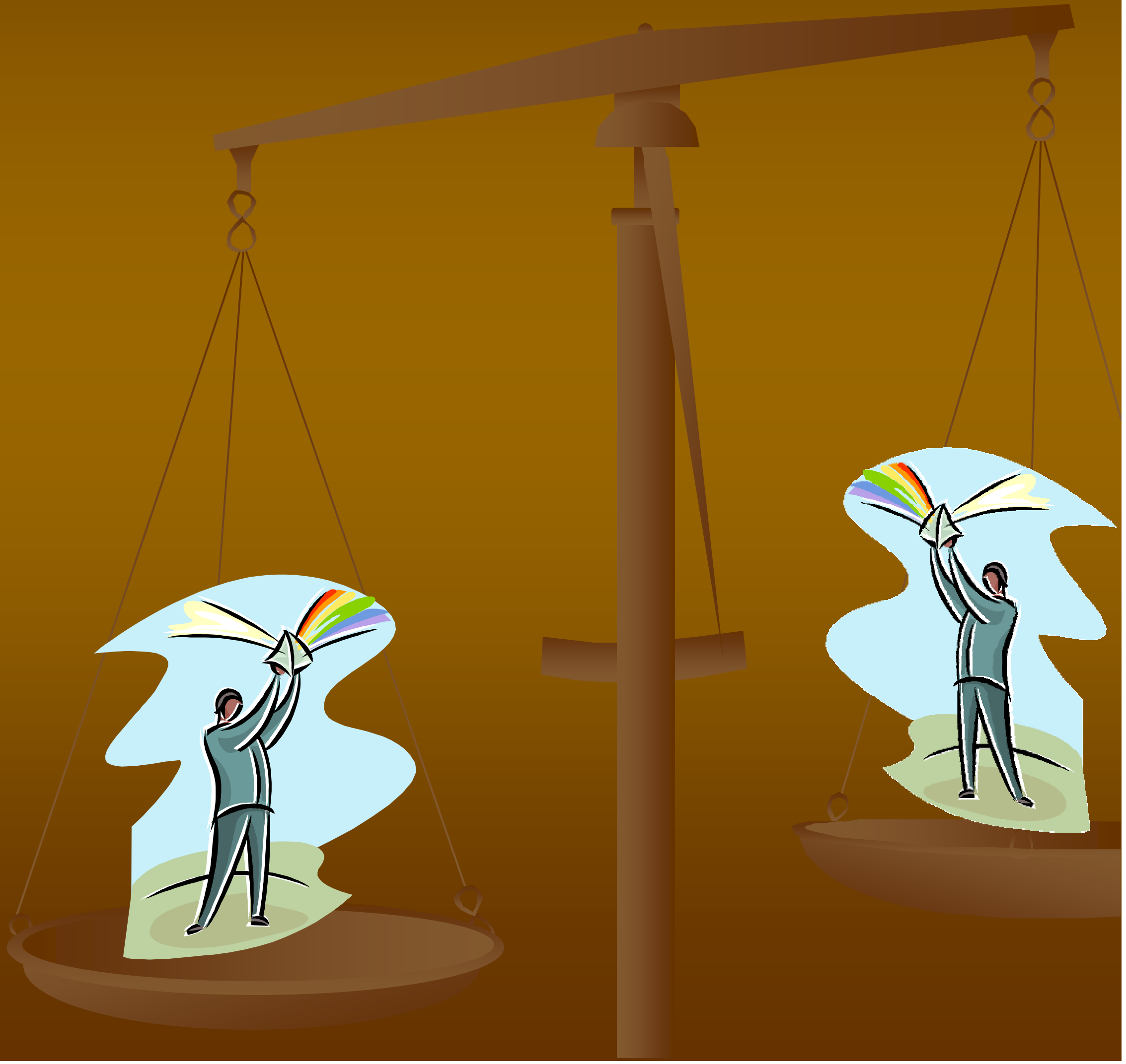
Meeting Purpose

- Regarding the standardized use of metrics in evaluation
  - Start framing the problem and laying out a path to proceed.



Spectrum of Evaluation

- Automatic Speech Recognition
- Machine Translation
- Optical Character Recognition
- Summarization
- Text to Speech
- ....



Automatic Speech Recognition

- Metrics
  - Word Error Rate (WER)
  - Additional Measures
    - Out of vocabulary rate
    - Task-based metrics

Information Retrieval

- Metrics
  - F-measure - the harmonic mean of precision and recall
    - **$F = (B^2 + 1) P R / ((B^2 P) + R)$** where

      P = precision = correct system responses / all system responses

      R = recall = correct system responses / all correct reference responses

      B = beta factor = provides a mean to control the importance of recall over precision
  - Additional Measures
    - Fallout – number of non-relevant responses / all non-relevant reference responses (related to, but not directly calculable from precision / recall)
    - False positives – items that are identified as correct responses that are not correct responses (= 1 – Precision)
    - False negatives – correct responses not identified (= 1 – Recall)
- Relevant Programs/Conferences
  - TIPSTER
  - TREC
  - NTCIR

Information Extraction

- Metrics
  - F-measure - the harmonic mean of precision and recall
    - F = $(B^2 + 1) P R / ((B^2 P) + R)$ where

      P = precision = correct system responses / all system responses

      R = recall = correct system responses / all correct reference responses

      B = beta factor – provides a mean to control the importance of recall over precision
  - Additional Measures
    - False positives – items that are identified as correct responses that are not correct responses (= 1 – Precision)
    - False negatives – correct responses not identified (= 1 – Recall)
  - Issues:
    - Classes of Entities
    - Annotation Standards for Development of Ground Truth
- Relevant Programs/Conferences
  - TIPSTER
  - MUC
  - MET
  - TIDES
  - ACE

Question Answering

- Metrics
  - F-measure - the harmonic mean of precision and recall
    - $F = (B^2 + 1) P R / ((B^2 P) + R)$ where

      P = precision = correct system responses / all system responses

      R = recall = correct system responses / all correct reference responses

      B = beta factor – provides a mean to control the importance of recall over precision
  - Additional Measures
    - False positives – items that are identified as correct responses that are not correct responses (= 1 – Precision)
    - False negatives – correct responses not identified (= 1 – Recall)
- Relevant Programs/Conferences
  - ARDA
  - NTCIR

Optical Character Recognition

- Metrics
  - UNLV ISRI Analytic Tools
    - Character accuracy
    - Marked character efficiency
    - Word accuracy
    - Non-stopword accuracy
    - Phrase accuracy
    - Cost of correcting automatic zoning errors
  - UMD's Multi-Lingual OCR Evaluation Tools (based on the UNLV's Comparison Tool)
  - Now … PAWs … more…?
- Some Relevant Programs/Conferences
  - ISRI's Annual Test of OCR Accuracy

Information Visualization

- Metrics
  - ??
- Relevant Programs/Conferences
  - ?

Other HLT

- Translation Memory
- Language Identification
- Transliteration
- Proper Name Matching
- Automatic Speech Recognition
- Text-to-Speech
- Audio Hotspotting
- ….

Another Question for this Workshop (#1)

- Which Human Language Technologies do we intend?
  - Just the ones represented here today?
  - Others, with as much inclusivity as possible?
  - ...?

Act 2: Rising Action (… or The Plot Thickens)

Machine Translation Evaluation: History

- Brand new field:
  - ~~2001, Kishore Papineni et al. introduce BLEU~~

    BLEU is interesting, but it isn't the whole story
- DARPA 1993 – 1994 MT Evaluation Campaign
  - Fluency, Adequacy, Informativeness
  - Task-based Evaluation (Task (error) tolerance)
- EAGLES / ISLE
- English → Russian → English

  "The spirit is willing but the flesh is weak" →
  "The wine is good but the meat is spoiled"
- English→ Chinese → French → English

  "Out of sight, out of mind" → "Invisible, Insane"
- Arabic → English

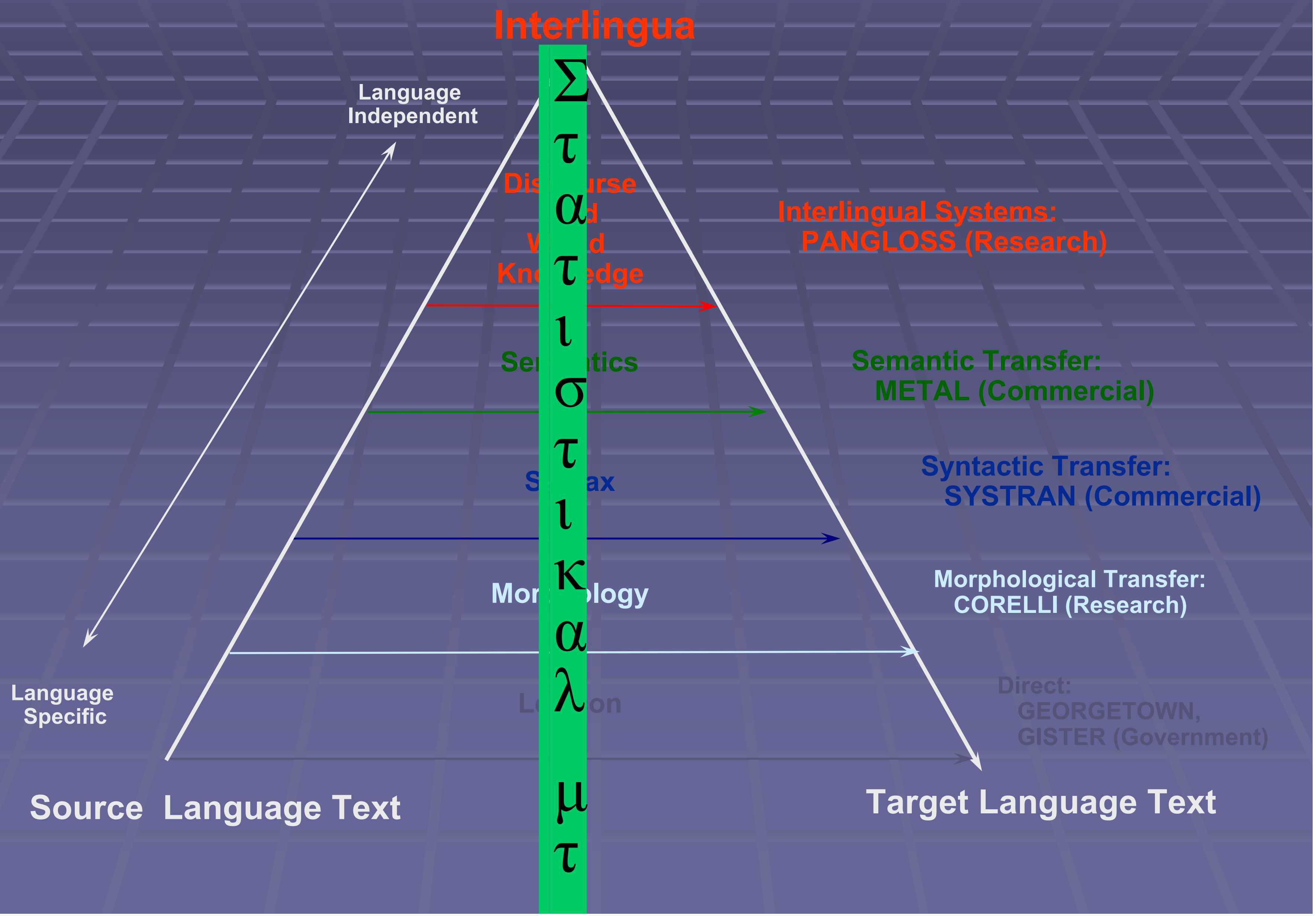
  ⟵ (2004/17/10) عنان : العالم ليس أكثر أمنا بعد الحرب على العراق

  Annan: the world not more secure after war to Iraq (10/17/2004)
- MT seeks to *emulate* human translators *for specific purposes*

Machine Translation

- Metrics
  - DARPA
    - Adequacy (Fidelity)
    - Informativeness (Fidelity)
    - Fluency (Intelligibility)
  - Task-based
    - Filtering
    - Detection
    - Triage
    - Extraction
    - Gisting (Summarization)
  - DLPT
  - PLATO
    - Clarity
    - Coherence
    - Syntax
    - Morphology
    - Unstranslated words
    - Domain terms
    - Proper Names
    - Adequacy (DARPA-style)
  - NEE
    - People
    - Organizations
    - Locations
    - Dates/Times
    - Money/Percentages

Levels of Knowledge



Interlingua

Language Independent

Language Specific

Dis urse
d
W d
Kn edge

Ser tics

S ax

Mor logy

Le on

Σ τ α τ ι σ τ ι κ α λ μ τ

Interlingual Systems:
PANGLOSS (Research)

Semantic Transfer:
METAL (Commercial)

Syntactic Transfer:
SYSTRAN (Commercial)

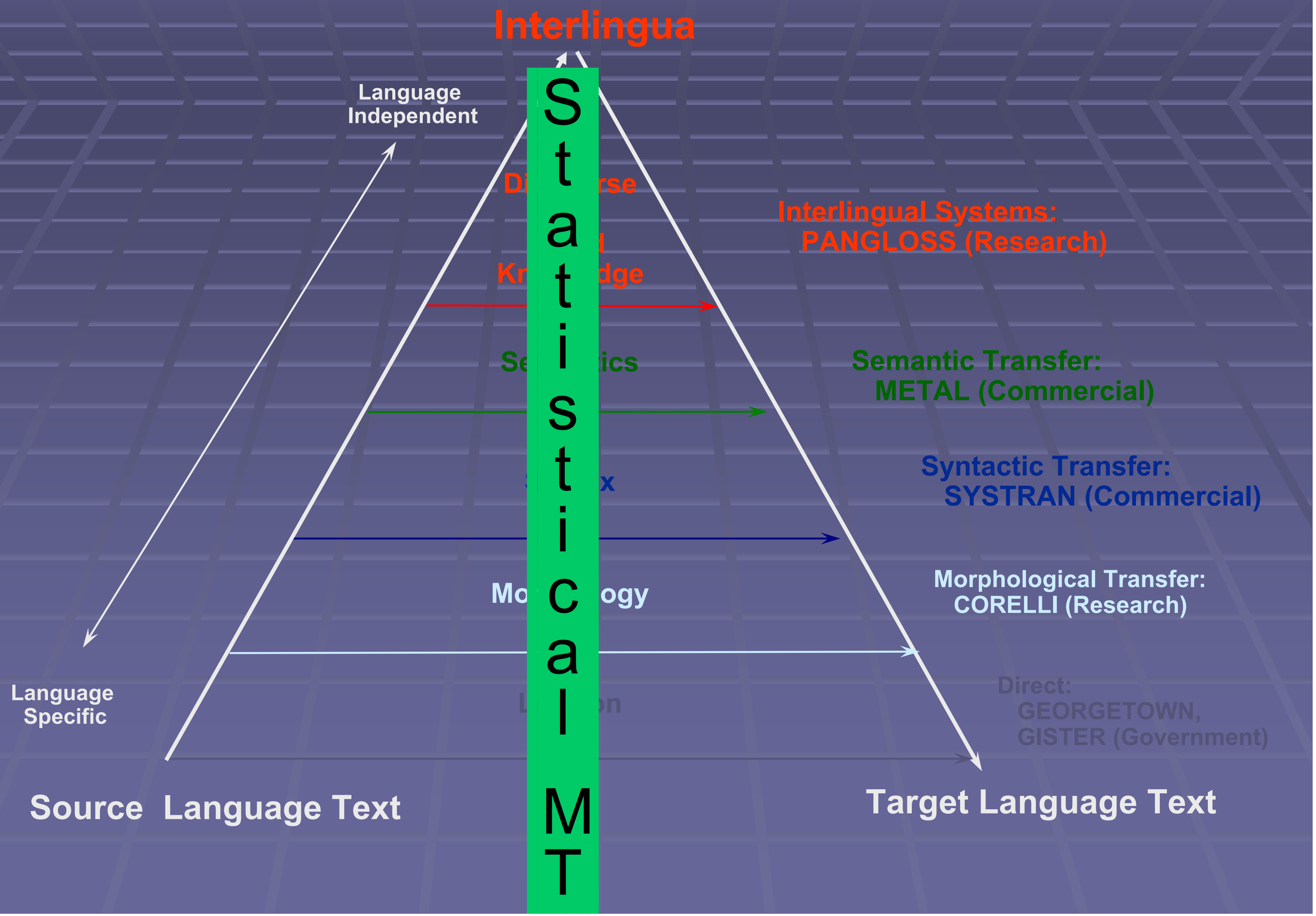
Morphological Transfer:
CORELLI (Research)

Direct:
GEORGETOWN,
GISTER (Government)

Source Language Text

Target Language Text

Levels of Knowledge



Machine Translation

- Metrics
  - DARPA
    - Adequacy (Fidelity)
    - Informativeness (Fidelity)
    - Fluency (Intelligibility)
  - Task-based
    - Filtering
    - Detection
    - Triage
    - Extraction
    - Gisting (Summarization)
  - **DLPT**
  - BLEU
  - NIST
  - ROUGE
  - PARIS
  - Edit Distance
  - D-Score
  - X-Score
  - **PLATO**
    - **Clarity**
    - **Coherence**
    - **Syntax**
    - **Morphology**
    - **Unstranslated words**
    - **Domain terms**
    - **Proper Names**
    - **Adequacy (DARPA-style)**
- Relevant Programs/Conferences
  - DARPA
  - FIDUL
  - TIDES
  - GALE
  - ELDA / ELRA??
  - **EAGLES**
    - **→ ISLE**
    - **→ FEMTI**
  - **NEE**
    - **People, Organizations, Locations**
    - **Dates/Times, Money/Percentages**

What Makes a Good Evaluation?

- Objective – gives unbiased results
- Replicable – gives same results for same inputs
- Diagnostic – can give information about system improvement
- Cost-efficient – does not require extensive resources to repeat
- Understandable – results are meaningful in some way to appropriate people

Framework for Evaluation: EAGLES 7-Step Recipe ➔ISLE (➔ FEMTI)

1. Define purpose of evaluation – why doing the evaluation
2. Elaborate a task model – what tasks are to be performed with the data
3. Define top-level quality characteristics
4. Produce detailed system requirements
5. Define metrics to measure requirements
6. Define technique to measure metrics
7. Carry out and interpret evaluation

PLATO:

Predictive Linguistic Assessments of *machine* Translation Output

Background

- Historical roots in DARPA evaluations of 1990s and subsequent work at FIDUL.
- Current activity emerged from a series of workshops on international standards for evaluating MT
  - ISLE – International Standards for Language Engineering
  - FEMTI – Framework for Evaluating MT in ISLE
- MT Summit 01, LREC, LREC Workshop ‘02
  - Distillation of seven linguistic tests for MT
  - Applications: similar SL/TL, SL/TL with greater divergence
- Results: Assessments appeared to rank systems

Relation to other work in MTE

- Automated MTE
  - BLEU (Papineni et al 2001)
  - BLEU + NEE (Papineni et al 2002)

- Task-based MTE
  - *Good Applications for Crummy MT* (Church and Hovy 1993)
    - EAGLES, ISLE, FEMTI
  - DARPA (White, Taylor, Doyon, others)
  - Reading comprehension / question answering (Jones et al)
  - CASL (Weinberg et al)

- PLATO
  - Relate linguistic signature of MT output to tasks
  - First necessary to determine quality of the metrics

Research Program Goals: *Linguistic Signature of MT Output*

- Develop a set of linguistic assessments for MT which, when applied to output, serve to predict the tasks which MT users can perform effectively on the output

- Through phased experimentation, establish:
  - **reliability and replicability of assessments**
  - correlations with automated measures
  - effect of varying input complexity/genre/medium
  - contribution of task performer experience/expertise

  - Automation of assessments
  - Automated determination of task suitability of MT systems

Linguistic Assessments

- Clarity
- Coherence
- Syntax
- Morphology
- Untranslated words
- Domain terms
- Names
- Adequacy (à la DARPA - added in most recent evaluation phase)

Approach

- Hire many assessors
  - Do they agree in their assessments?
  - Can we model a task with the scores?
- Teach assessments
- Develop guidelines for assessments
- *Measure Agreement*
- Refine assessments and guidelines
- Re-Measure Agreement
- Repeat to determine improvement in metrics’ reliability

Inter-Assessor Agreement

- Joint agreement (weighted)

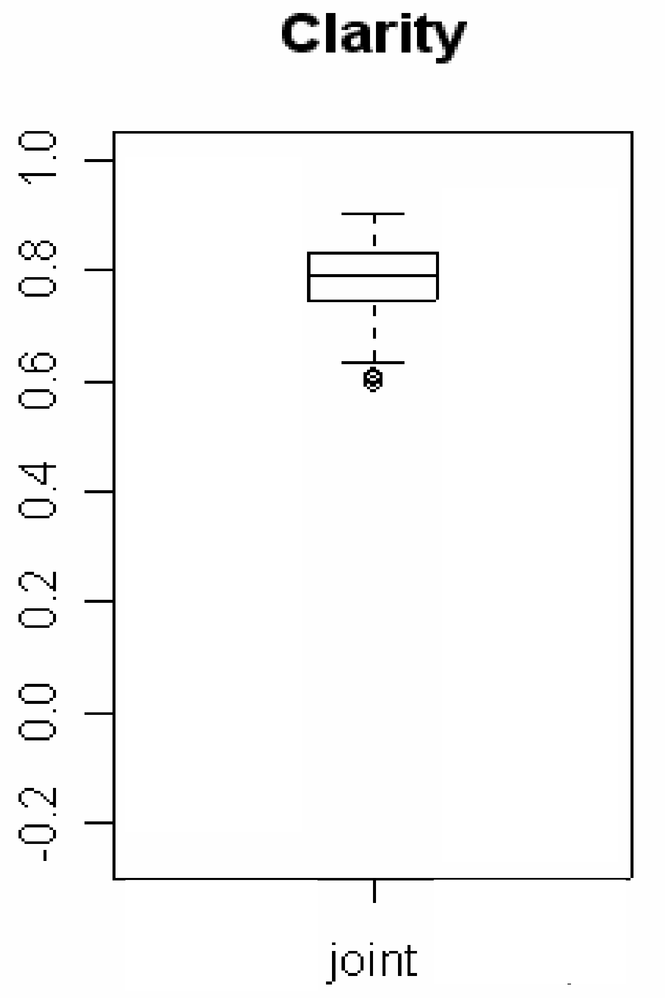
$$\sum_{i=\min}^{\max}\sum_{j=\min}^{\max}\omega_{i,j}p_{i,j}$$

$$\omega_{i,j}=\frac{|i-j|}{\max-\min}$$

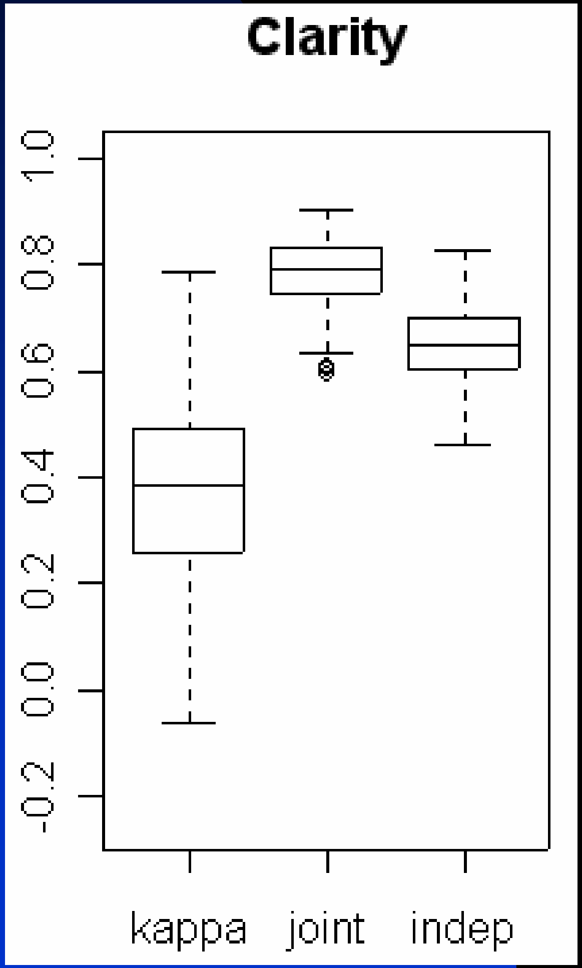
$p_{i,j}\equiv$ proportion of observations in the cell at row i, column j

$p_{i,\bullet}\equiv$ proportion of observations in row i

$p_{\bullet,j}\equiv$ proportion of observations in column j



Goal: Metrics with High Reliability



Kappa (artificially) low due to high independent probability of agreement.

Dependent on affinity of single assessors for particular ratings

Dependent on homogeneity or variability in texts being assessed

Methods of addressing

- lower independent
- raise joint
- statistic

Another [Side] Question for this Workshop (#2a)

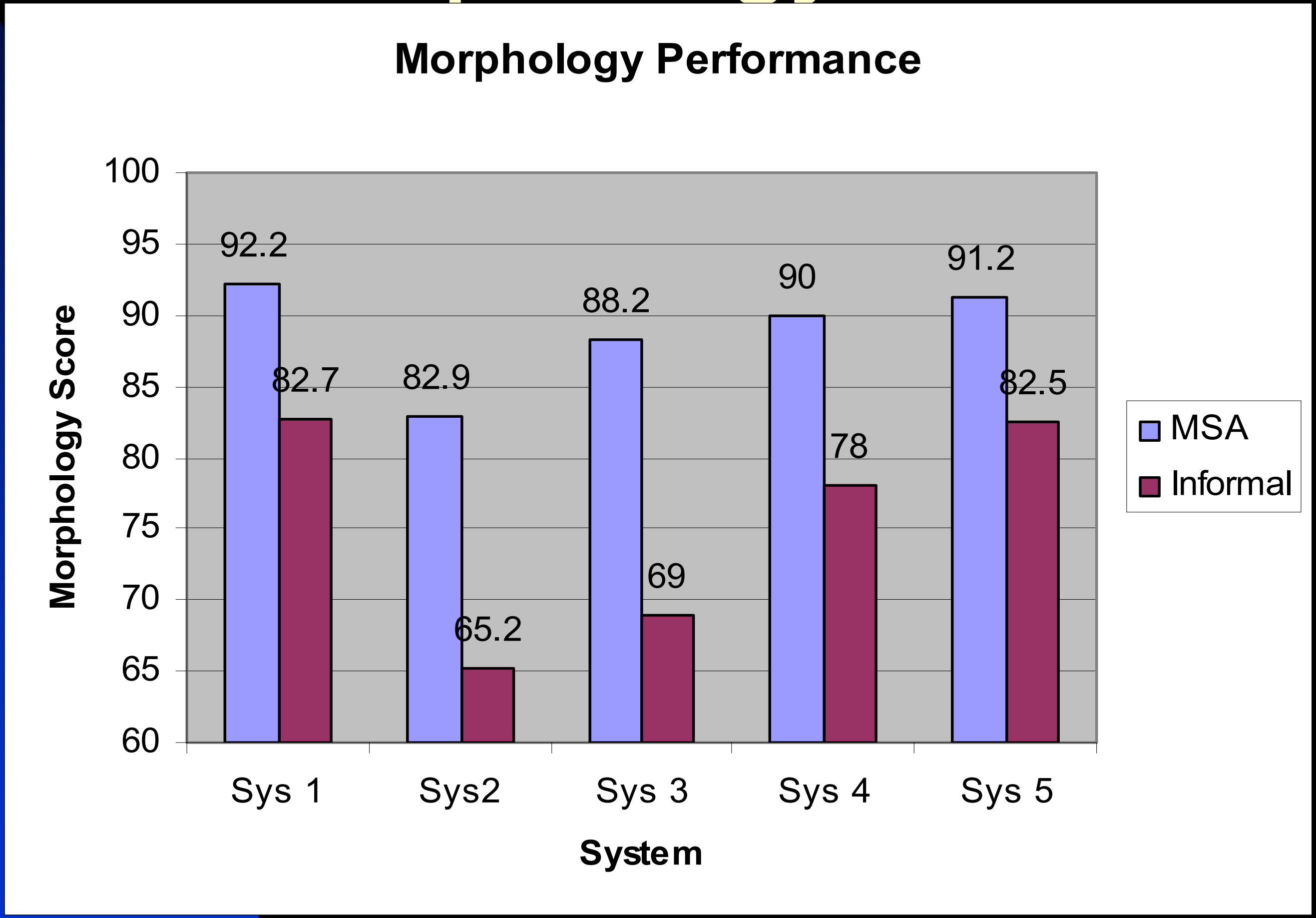
- Is kappa **the** test statistic that we should be using to test interrater agreement (when the chosen evaluation paradigm rests crucially on creation of ground truth data by human annotators and on the quality of that ground truth data)?
  - If yes, how should it be modified for cases in which it isn’t a perfect fit?
    - Think about BLEU, as an extreme case
  - If no, what other statistics / quality checks should be developed?

Goal:

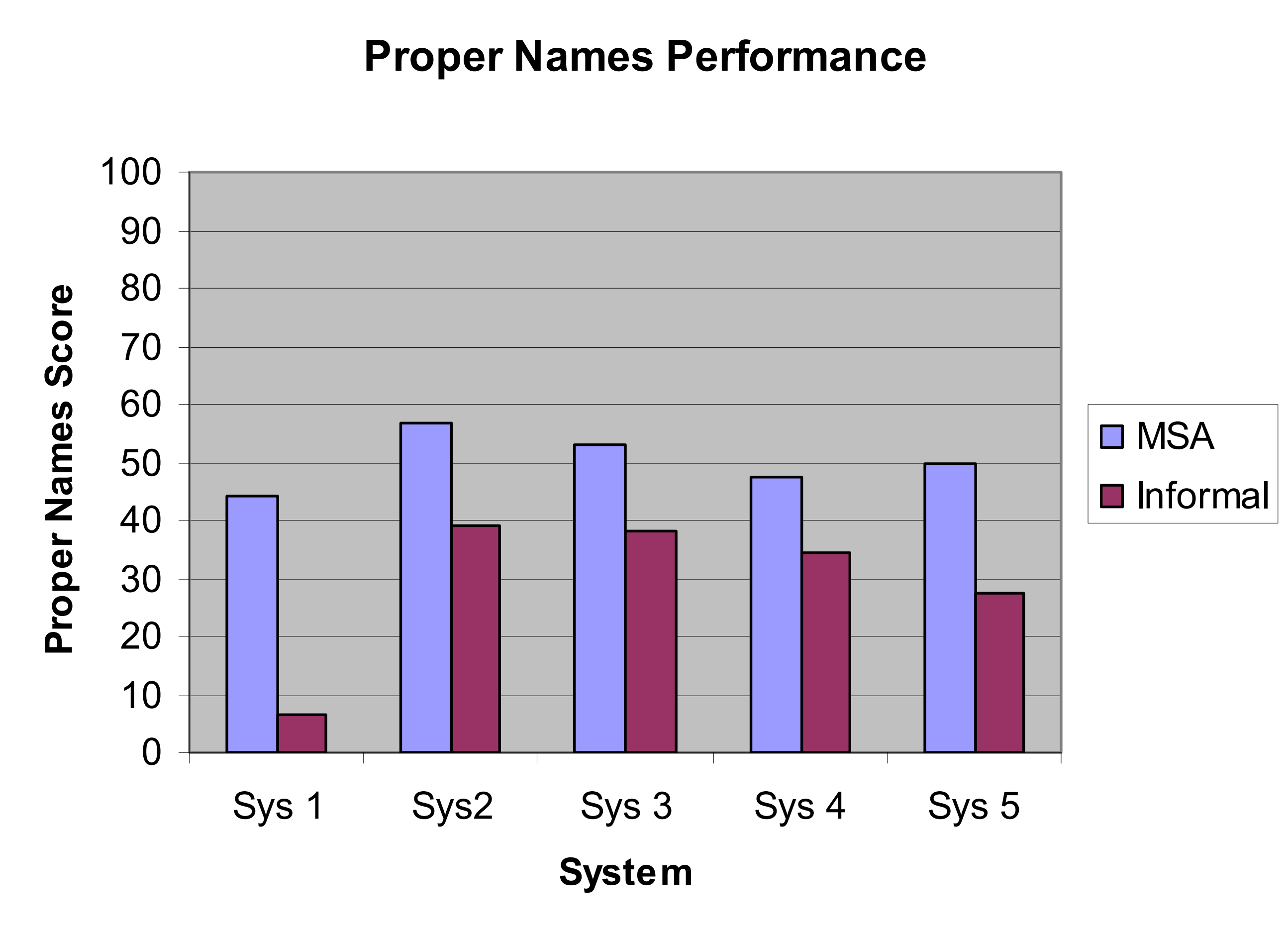
Linguistically-Based Metrics with High Reliability

- Interpretable
- Relate to Utility of Output

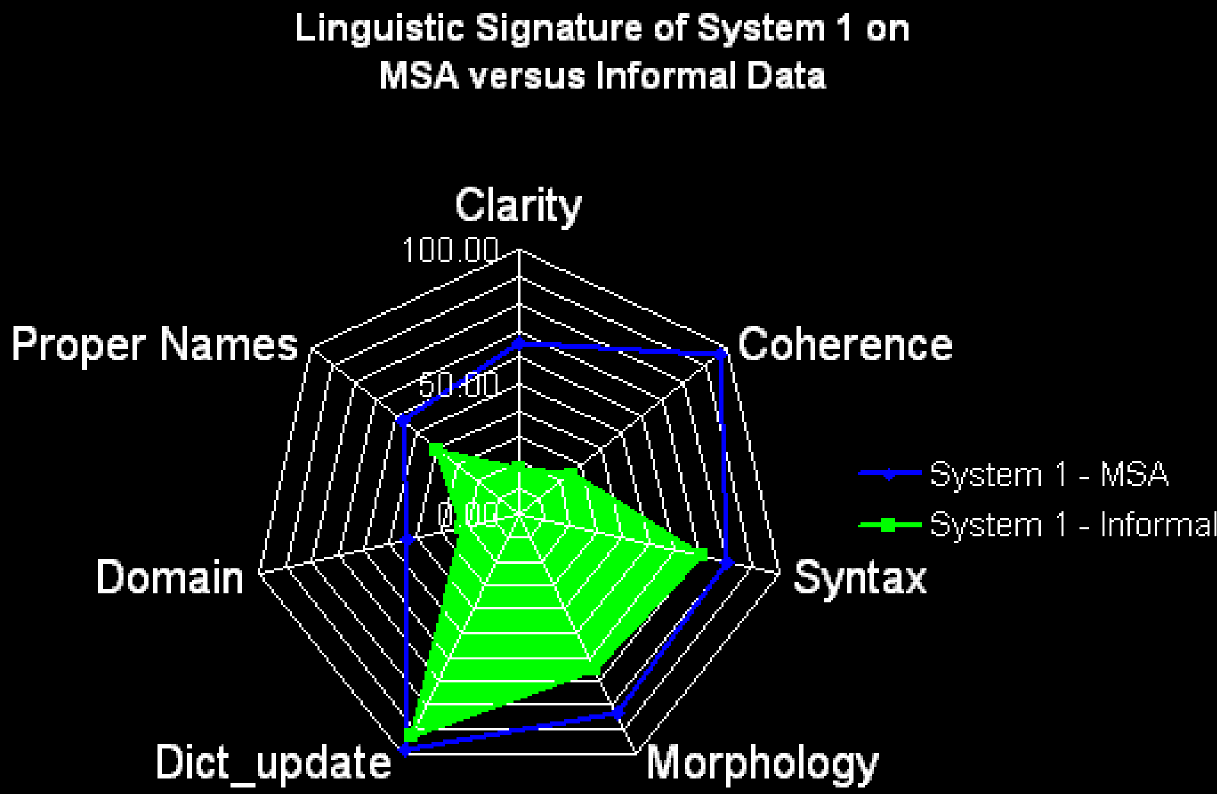
PLATO-O Arabic MT Assessment: Morphology



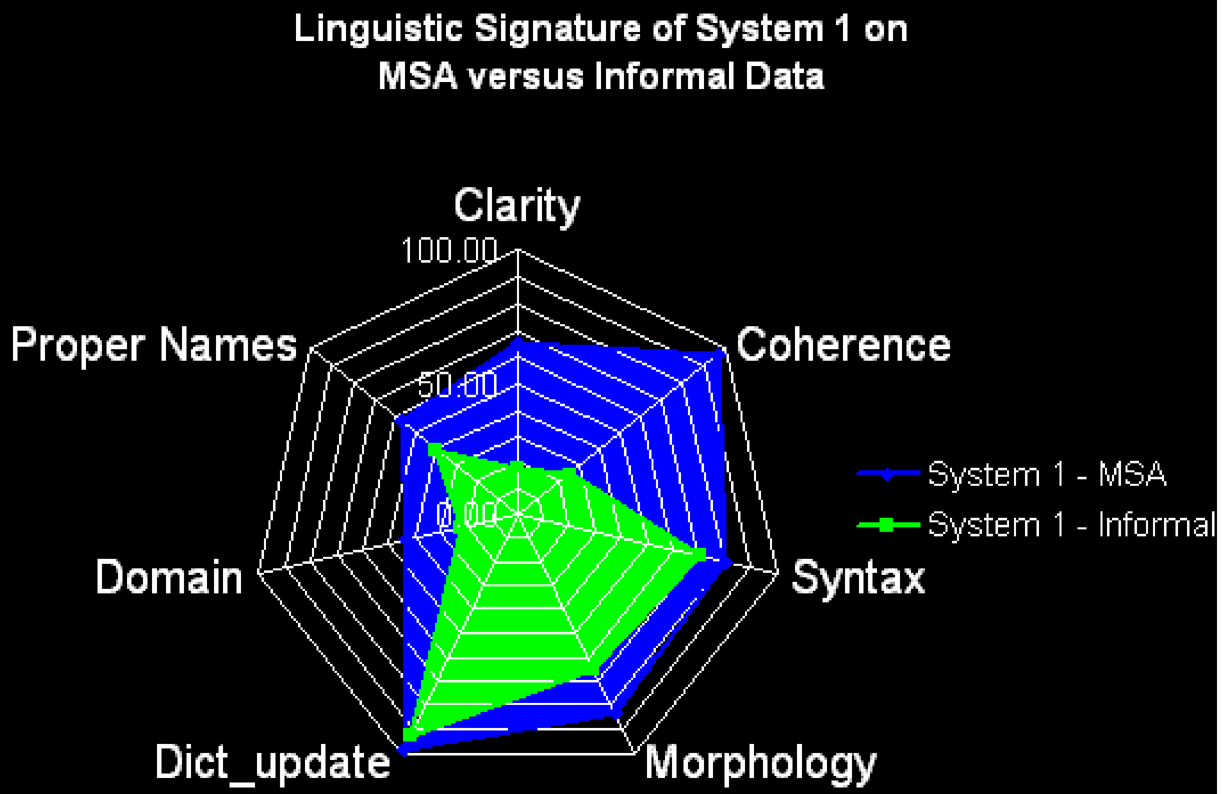
PLATO-O Arabic MT Assessment: Proper Names



PLATO-O Assessment: Arabic MT: MSA vs Informal



PLATO-O Assessment: Arabic MT: MSA vs Informal



Where from here?

- Correlation of Linguistic Signatures with Tasks
- Correlation of Assessment scores with Automated Metrics
- PLATO Operational Evaluation
- PLATO Evaluation of MT in Embedded Contexts:
    - Degradation from preprocessing
        - OCR+MT
    - Appropriateness for downstream processing
        - MT + IE
- Refinement of metrics

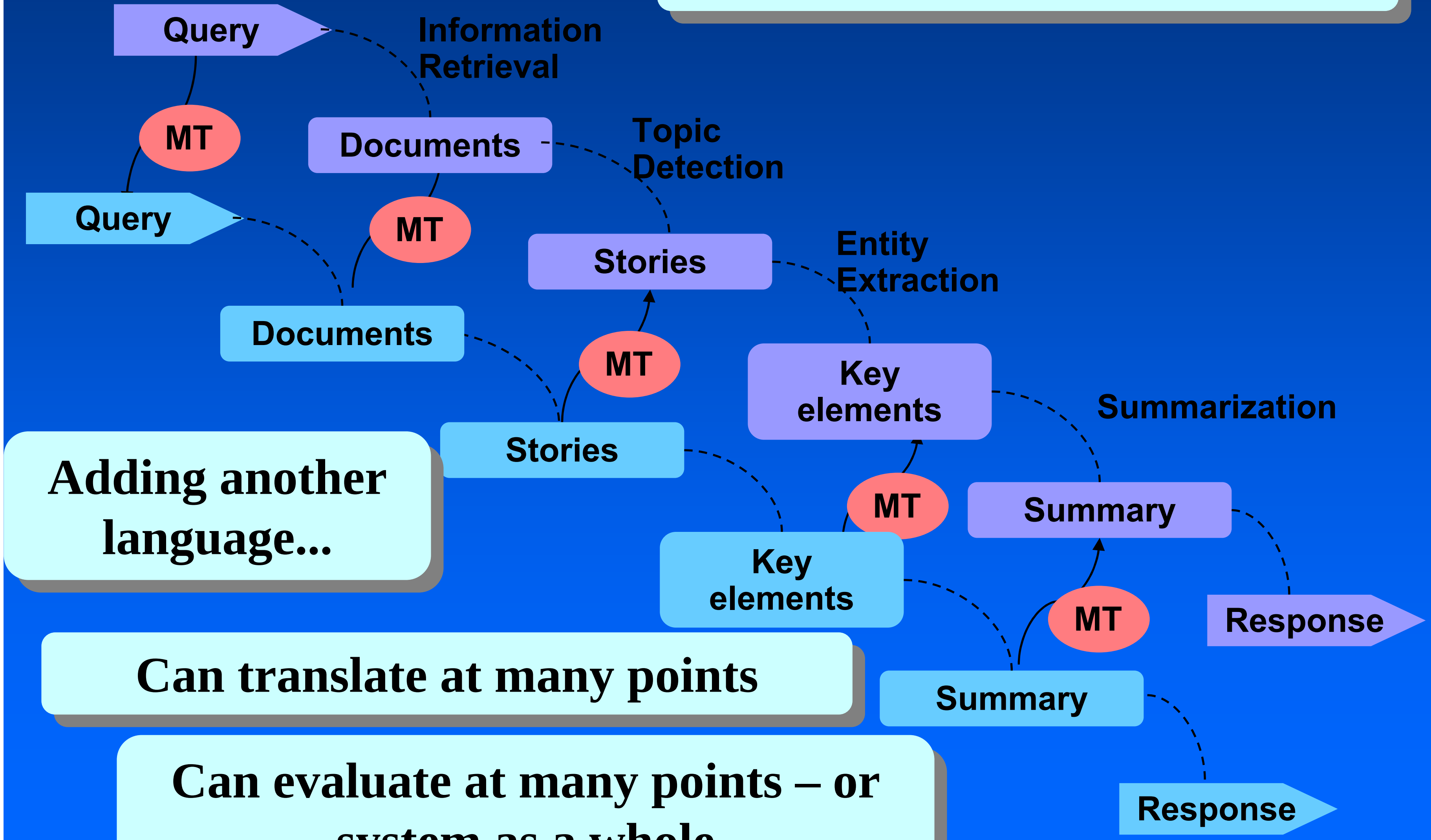
Another Question for this Workshop (#2b)

- Is this an example of a useful paradigm for doing research in HLT evaluation (roughly outlined as the following)?
  - Identify tasks of importance
  - Identify features important to those tasks
  - Define metrics to measure system performance on these features
  - Determine actual correlation between metrics and suitability of system output to task(s)
  - If metrics are prohibitively expensive to perform on an ongoing basis, search for automated metrics that correlate with human-based metrics and with task performance.

Act 3: Cliffhanger

Putting Components Together

Monolingual query-to-answer



Adding another language...

Can translate at many points

Can evaluate at many points – or system as a whole

Information Extraction Tool Suites

- From component-level evaluation to end-to-end systems evaluation
    - Isolated component-level evaluation
    - Embedded component-level evaluation
    - End-to-end system evaluation
- Metrics
    - Usability
    - Performance/Functionality
        - Black box
        - Glass box
- Relevant Programs/Conferences
    - ?

Maghi King: « Relevant to research evaluation »

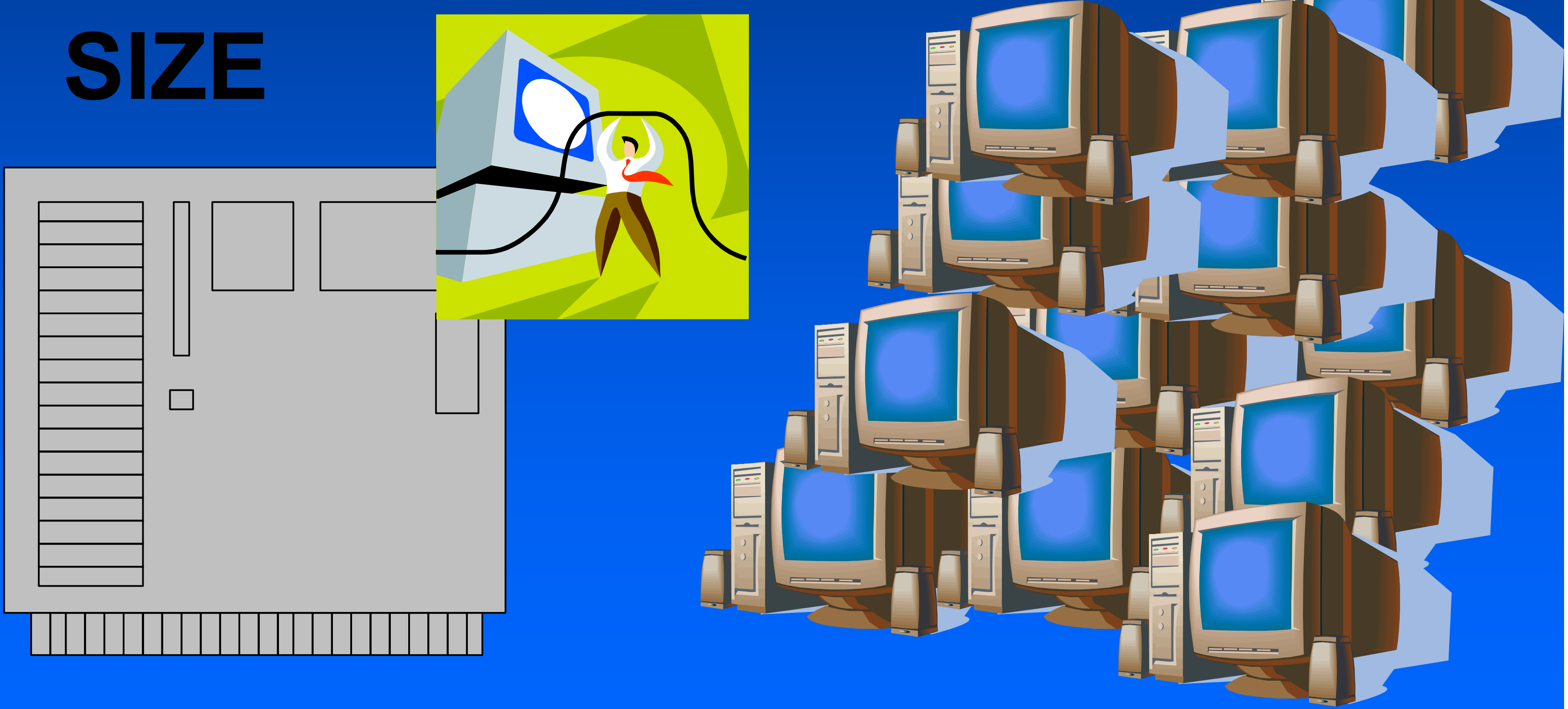
- The ISO quality characteristics
  - Functionality
  - Reliability
  - Usability ?
  - Efficiency
  - Maintainability
  - Portability ?

Should the R&D community be worrying about anything besides quality?

SPEED

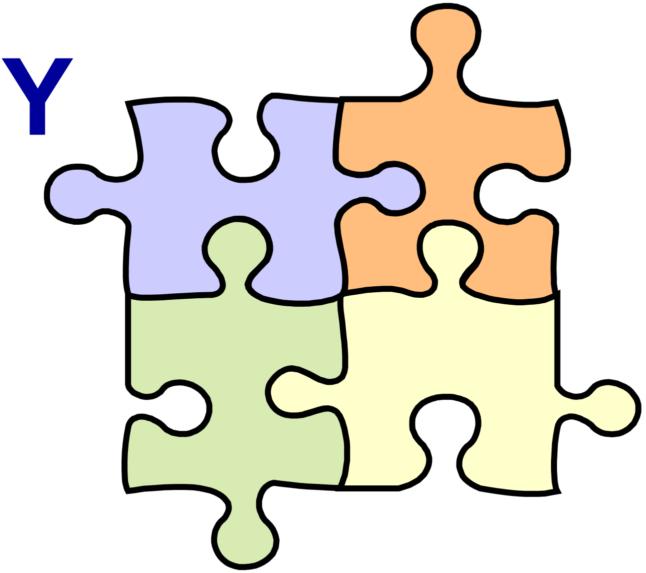
of throughput

SIZE

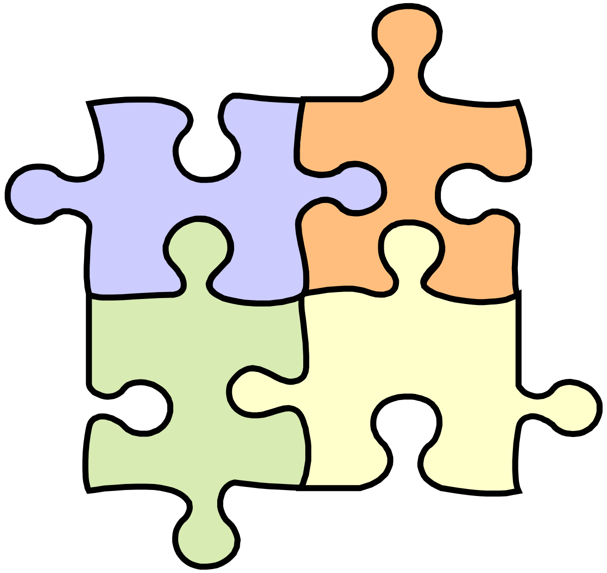


And…

CONFIGURABILITY



EMBEDABILITY



Should the R&D community be worrying about anything besides quality?

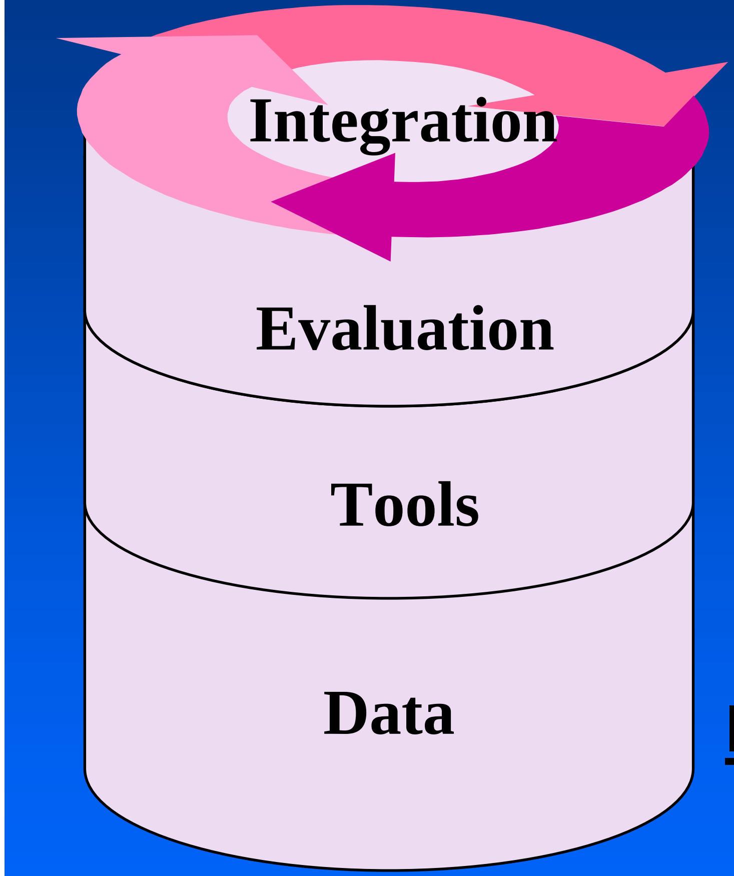
- Speed

- Size of deployment (platform):
  - room-size
  - mini, PC, handheld
  - server farm....

- Configurability: user dictionaries, domain dictionaries, speed/quality tradeoffs, etc.

- Embedability: APIs (ease of use, granularity)

The Underlying Drivers of Success



Data: Evaluation (ground truth and other) data, training data, usability data drive progress

Modular approach

Tools support data creation

Modules provide reusable component-ware

Metrics-based evaluation:

What works and how well?

… and to what end?

Integration and embedding

The whole is greater than the sum of its parts!

Must be evaluated as such: component-level and system-level evaluation

A Final Question for this Workshop (#3)

- Is it possible for HLT Evaluation to serve the multiple masters it is beholden to?
  - System selection
    - Stand-alone systems
      - Component-level evaluation
    - Embedded-systems
      - Component-level and/or system-level evaluation
  - Research
    - Progress in basic capabilities and functionality
- Can we do this and still conduct principled research in (useful) evaluation methodologies?