Odds of Successful Transfer of Low-Level Concepts:

A Metric for Speech-to-Speech Machine Translation

Gregory A. Sanders, Sébastien Bronsart,
Sherri Condon, Craig Schlenoff

MITRE

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

The TRANSTAC Goal

Enable U.S. personnel speaking only English to communicate with civilian populations speaking only other languages

The TRANSTAC Program

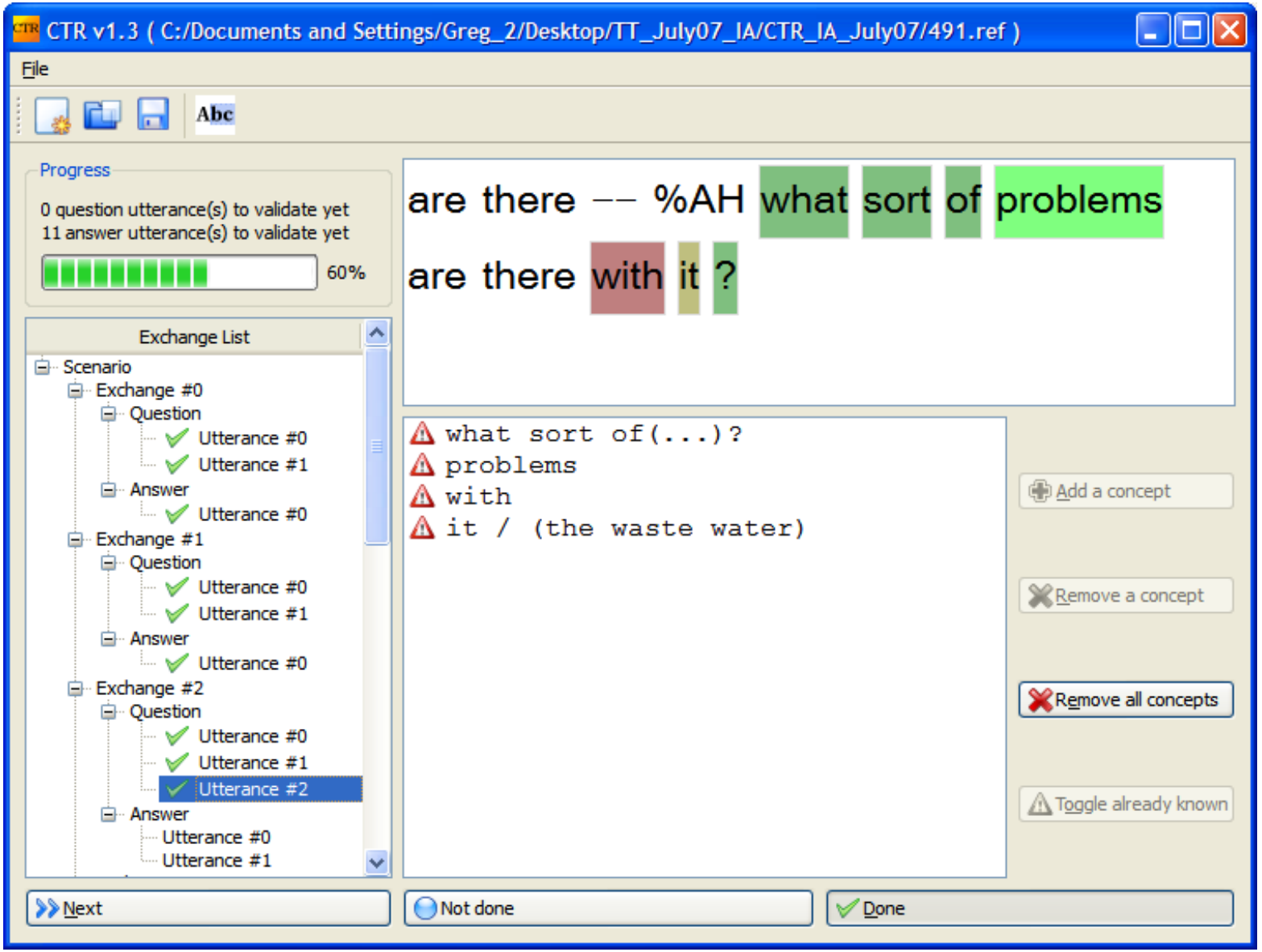
Spoken Language Communication and Translation System for Tactical Use

- Bidirectional Speech-to-Speech Machine Translation
- Laptop or hand-held platform
- Free-form input, but in known domains
  - Medical
  - Civil Affairs
    - Examples: Sewer, Water, Electricity, Trash
  - Military operations
    - Examples: Training, Joint ops, Vehicle checkpoint
- Program sponsored by DARPA
- System performance evaluated by NIST and MITRE

What are the Low-level Concepts?

- We defined the low-level concepts to consist of the source-language content words
  - Open-class words:
    - Nouns, Verbs, Adjectives, Adverbs
  - Important prepositions and quantifiers
  - Entire verb construction (e.g., “will have been thrown”) is one concept
- Speakers choose what to give prominence via expression as a content word
- Number of such elements is determined by the speaker
  - Count is not open-ended
  - Count is not highly subjective
- Low-level concepts annotated in the source-language transcript
  - Annotated by a native speaker
  - If utterance is disfluent, count only the concepts that a fluent rendition would include

CTR, in reference annotation mode



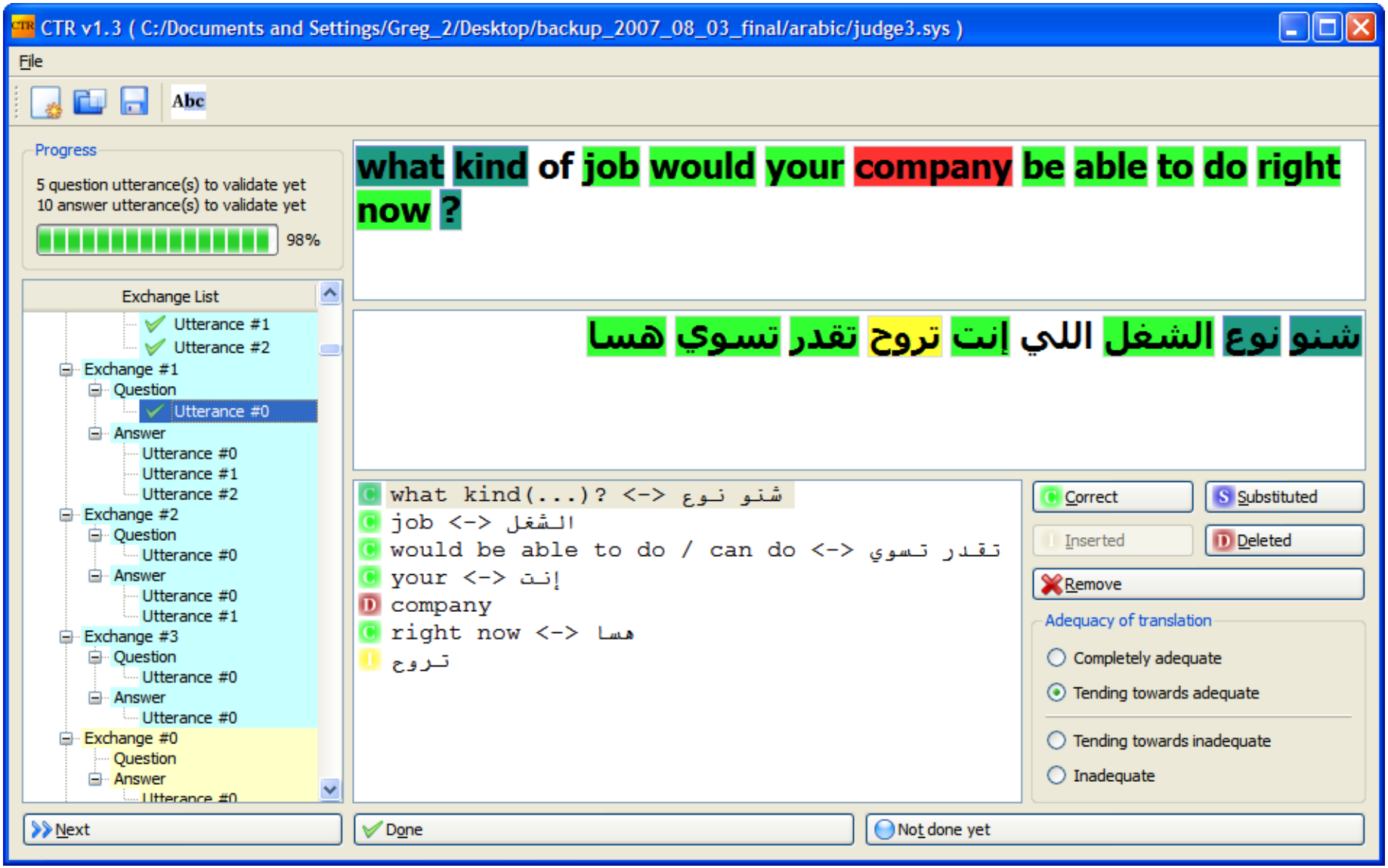
Scoring Successful Transfer

- Panel of bilingual judges who each score the MT output
  - Compare textual target-language MT output to annotated transcription of source-language utterance
  - Each low-level concept is scored:
    - Successfully transferred --- Correct
    - Deleted
    - Substituted
    - Inserted concepts are also identified by the judges
- Result stated as Odds of Successful Transfer of a low-level concept

  Odds(*correct*) =

  NumCorrect / (Deletions + Substitutions + Insertions)
- Progress across evaluations can be stated as an Odds Ratio

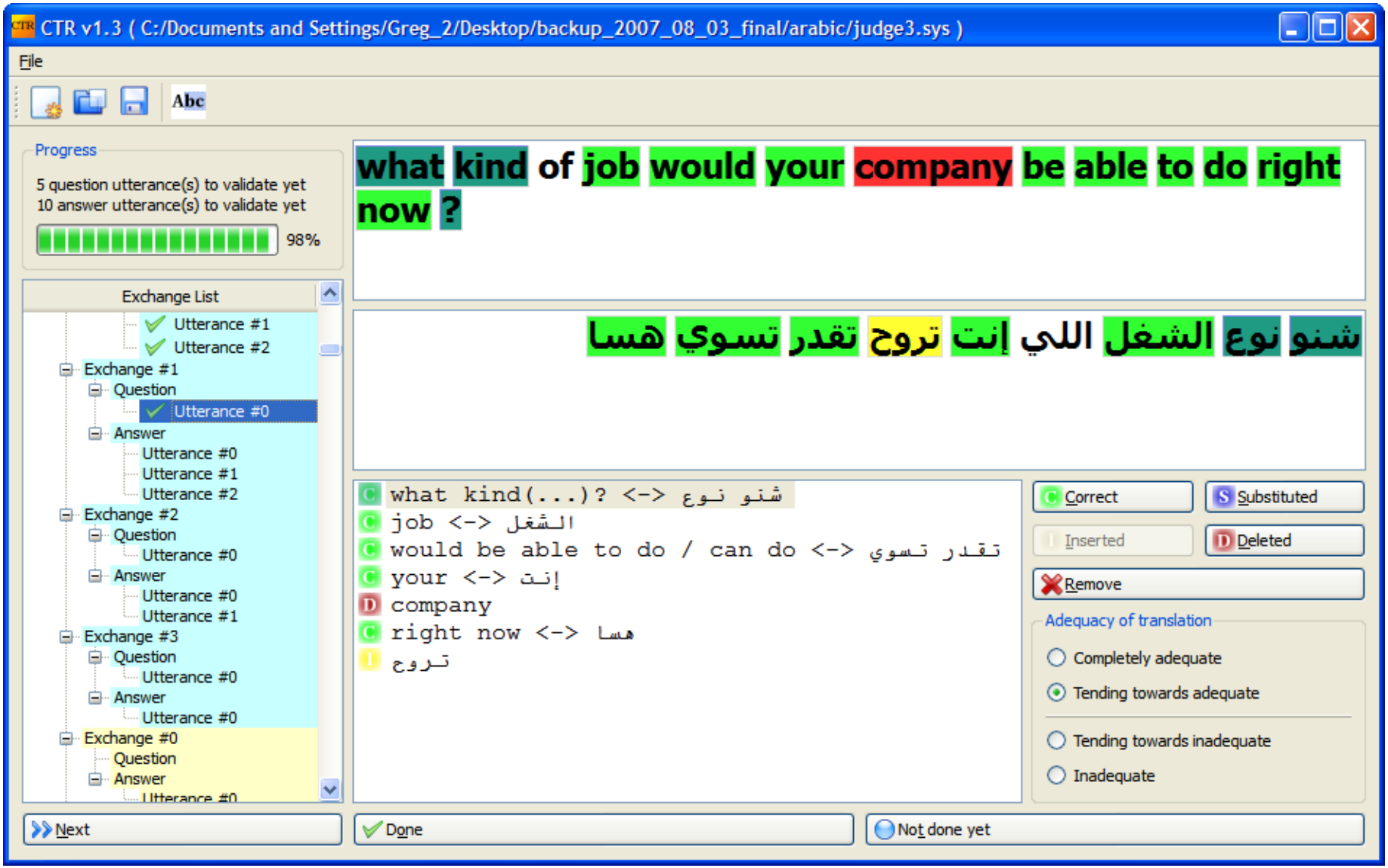
CTR in MT output scoring mode



Judgments of Semantic Adequacy

- We asked our bilingual judges to also give a single judgment of semantic adequacy for each utterance on a four-point scale
  - Completely adequate
  - Tending towards adequate
  - Tending towards inadequate
  - Inadequate
- Judges assigned this utterance-level score immediately after scoring the low-level concepts in the utterance
- We consider these judgments to be our *benchmark* score
  - We compare our other metrics to it

CTR in MT output scoring mode



Training the Judges for Semantic Adequacy

- We explained the intended use and purpose of the system
  - Asked judges to assign scores that reflect how well the translations would serve that purpose
- We gave the judges a substantial set of exemplars for each of the four possible scores
  - The exemplars were taken from a previous eval, and were utterances on which the (different) set of judges from that eval had a high level of agreement
- We had the judges discuss several example translations as a group
  - Made sure each judge was offering appropriate reasons for their choice of score --- made sure they understand the task
- For Arabic, we told the judges to favor translations into Iraqi dialect, not the standard written language (MSA or Fus'ha)

Converting Odds to Probability of Correct Transfer

NumCorrect / (Deletions + Substitutions + Insertions)

- Because we count insertions as errors, our odds calculation is not quite canonical P(*correct*) / (1 – P(*correct*))
- As P(*correct*) approaches 1.0, Odds(*correct*) approaches ∞
  - Typical automated MT metrics behave mathematically more like P(*correct*) than like Odds(*correct*)
  - Correlation with automated MT metrics calls for a statistic that behaves like P(*correct*), but with insertions taken into account
- Adjusted Probability Correct

$$AdjP(correct) = 1 - (1 / (Odds(correct) + 1))$$

Other metrics are also important

- Concepts vary in importance -- some concepts are crucial
  - Utt: There are new IEDs along the road from here to Fallujah.
  - MT: There are no IEDs along the road from here to Fallujah.
- Low-level concept transfer metric gives all concepts equal weight
  - Utterance-level human judgments of semantic adequacy weigh the crucial errors appropriately
- Low-level concept transfer metric does not consider fluency
  - Even badly fractured syntax may be given a pass
  - Many automated MT metrics (e.g., BLEU, METEOR) do effectively consider fluency, as do utterance-level human judgments

Other Metrics We Calculated

- Source-language ASR was scored with Word Error Rate
- MT was scored with several commonly used metrics
  - BLEU
  - METEOR
  - TER
  - HTER --- only completed for translations into English

Discussion of Results

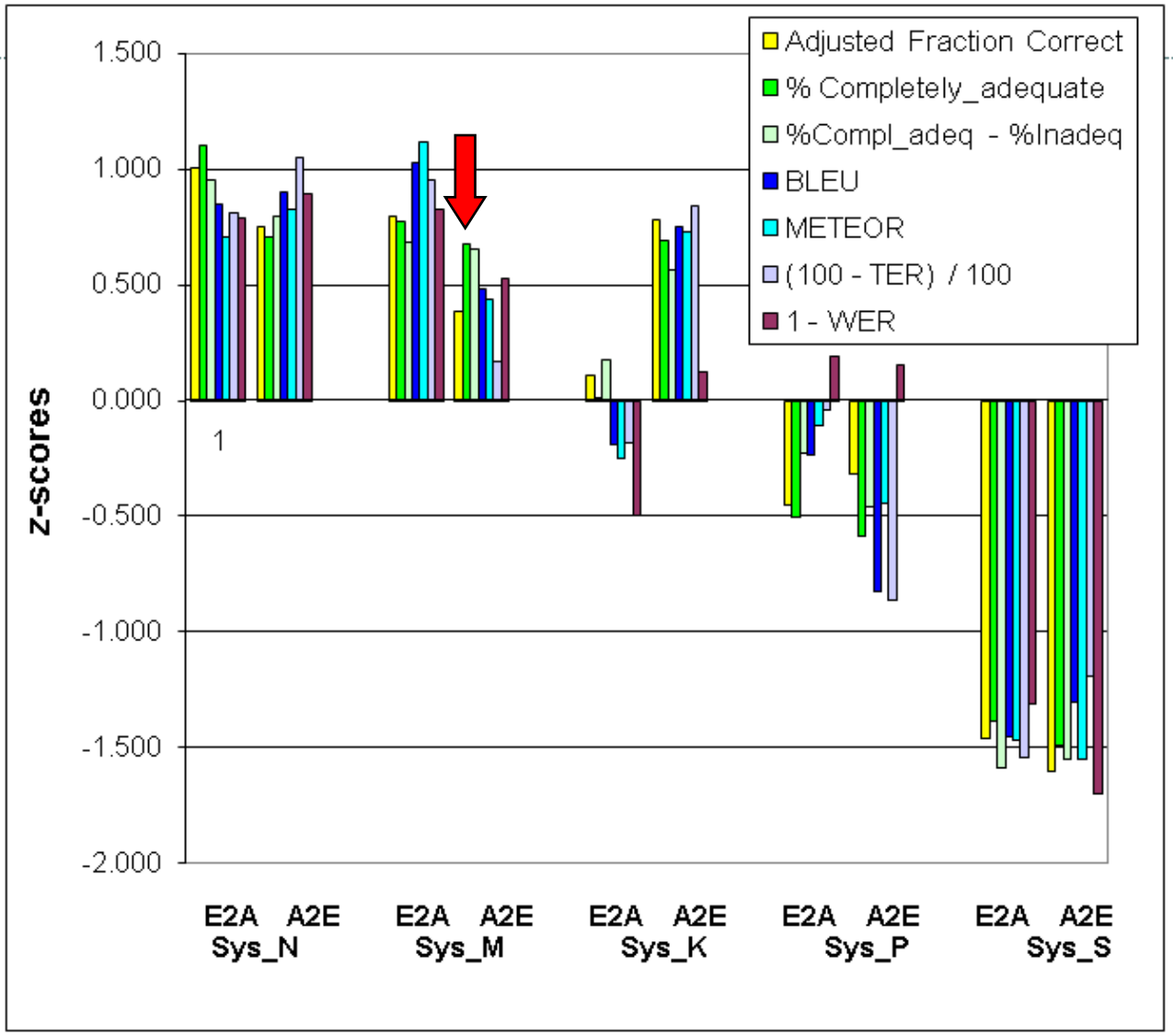
- Between January 2007 and July 2007 systems made large improvements in this metric
  - For English to Iraqi Arabic the median value over the five systems improved to 4.32 from 1.55 (an odds ratio of 2.79)
  - For Iraqi Arabic to English the median value improved to 3.15 from 2.46 (an odds ratio of 1.28)
- Scores on AdjP(*correct*) strongly correlated to the utterance-level judgments of semantic adequacy
  - Pooling all data for each system, Pearson correlation over the five systems
    - R = 0.997 for English to Iraqi Arabic
    - R = 0.978 for Iraqi Arabic to English
    - R = 0.997 for English to SurpriseLanguage
    - R = 0.960 for SurpriseLanguage to English

Comparing all the Metrics

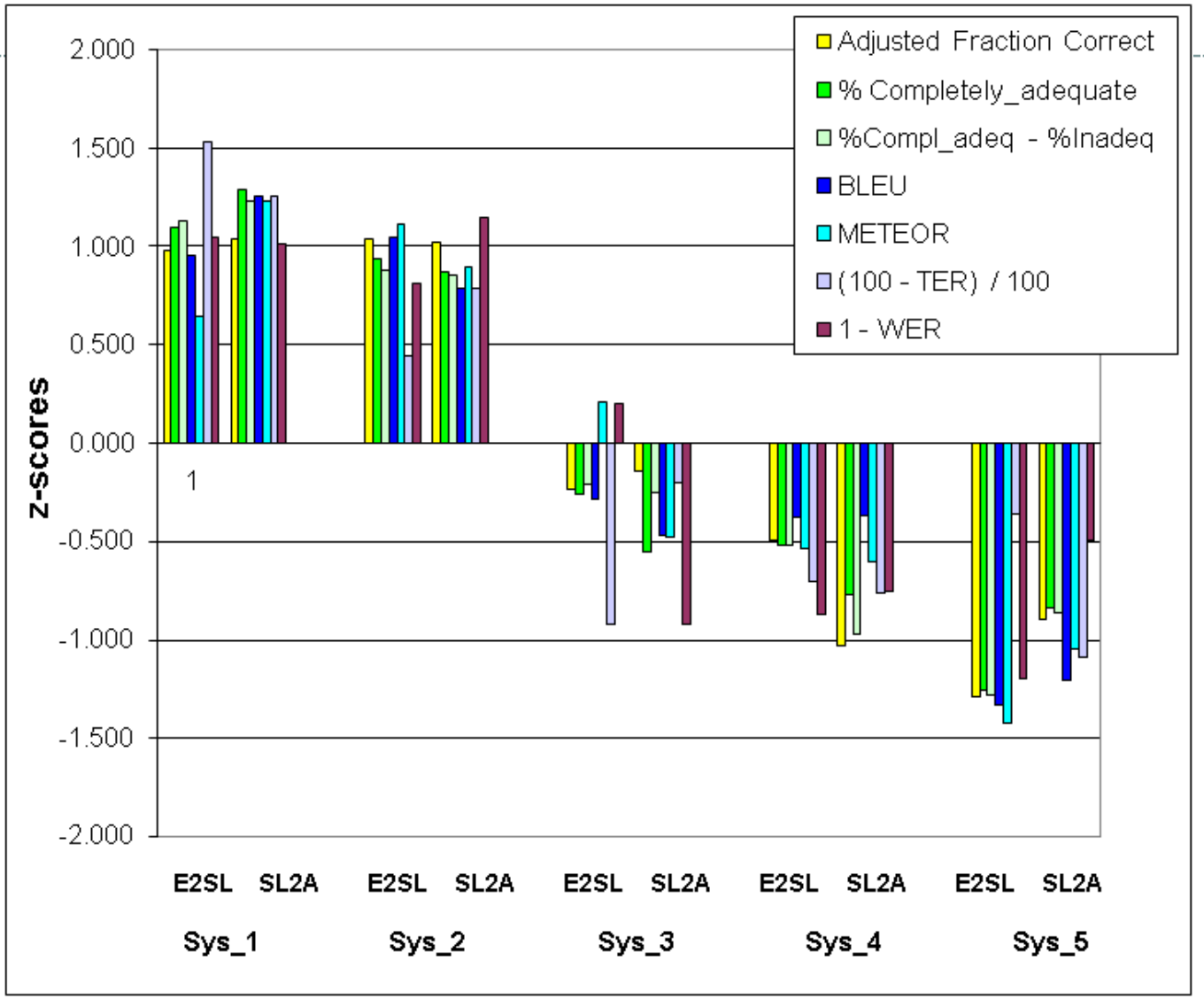
- For each language pair, separately, and each direction (to/from English) separately, we calculated mean and standard deviation, then converted all values to standard normal *z* statistics

- Result shown in the following synoptic overview graphs

Synoptic Overview for Arabic



Synoptic Overview for SurpriseLanguage



HTER for Iraqi Arabic to English

- HTER based on a human post-editing the MT output as necessary so that it has the correct meaning (fix the semantic errors)
  - HTER is a measure of the minumum number of edits necessary
- Key wrinkle in TER and HTER: a block move counts as one edit
  - Moving a string of any number of words by any distance
- Looking at HTER for each of the nine scenarios, for each of the four strongest systems (thus 4 x 9 = 36 data points)
  - Pearson correlation of HTER with AdjP(*correct*) is R = 0.905
  - Pearson correlation SemAdeq with AdjP(*correct*) is R = -0.833
- Omitting the hardest and easiest scenario to eliminate outlier effects (thus, 4 x 7 = 28 data points)
  - Pearson correlation of HTER with AdjP(*correct*) is R = 0.849
  - Pearson correlation SemAdeq with AdjP(*correct*) is R = -0.790

Inter-judge Agreement on Semantic Adequacy

- We had six judges for Arabic, and five for the surprise language
- Values of Cohen's kappa for pairwise inter-judge agreement, over the Arabic judges:
  - Exact match pairwise kappa range 0.178 to 0.435 (median 0.294)
    - Very low values -- not good
  - If we count the disagreements by just one level as being matches, then the pairwise kappa range is 0.508 to 0.805 (with median 0.611)
    - We regard this as an acceptable level of agreement
- For odds of successful transfer, there was fairly close agreement between the mean and median values over our set of judges
- Considering all this, we suggest that a reasonably large set of judges is necessary, as outlier judges are likely

Conclusions

- Odds of successful transfer of a low-level concept appears to be a relatively useful quantitative metric for information transfer
  - Strong correlation to human judgments of semantic adequacy
  - Strong correlation to the most common automated MT metrics, such as BLEU and METEOR
- The metric is labor-intensive
  - More useful for summative evaluation
- Training the judges carefully is important
  - Important to provide guidelines, with several examples of what counts as the same and what counts as different. Tricky tricky issues arise.
- Using a panel of ***several*** bilingual judges appears important
  - There were notably forgiving and harsh judges (outliers).
- Getting some judges to mark insertions Is difficult; this can bias results.

For Further Info

- Over time, various TRANSTAC papers, presentations, guidelines documents, and so forth, will appear in the web pages for the NIST Speech Group

  http://www.nist.gov/speech