Evaluating Lexical Coverage for Cross-Language Information Retrieval
Gina-Anne Levow and Douglas W. Oard, University of Maryland, USA
gina@umiacs.umd.edu oard@glue.umd.edu http://www.umiacs.umd.edu/labs/CLIP/

Nature of the Problem
• Translation lexicon coverage is important
  – Coverage gaps cause retrieval failures
• Existing measures have weaknesses
  – Term counts can't reveal topical differences
  – Computing retrieval effectiveness is expensive

Desirable Characteristics
• Insightful
  – Some degree of predictive validity
• Affordable
  – Easily computed without expensive resources
• Single-valued
  – Facilitates comparison of alternative resources

Results
• Bigger isn't always better
  – LDC has three times as many English words
• TF and TFIDF are both useful
  – TF is easier to compute
• Merged lexicons pose unique challenges
  – No measure predicted average precision well

Next Steps
• Evaluate small and specialized lexicons
  – Might show an advantage for TFIDF
• Estimate confidence intervals
  – How much change in a measure is significant?

More on Cross-Language Retrieval:
http://www.clis.umd.edu/dlrg/clir/