AMTA-06

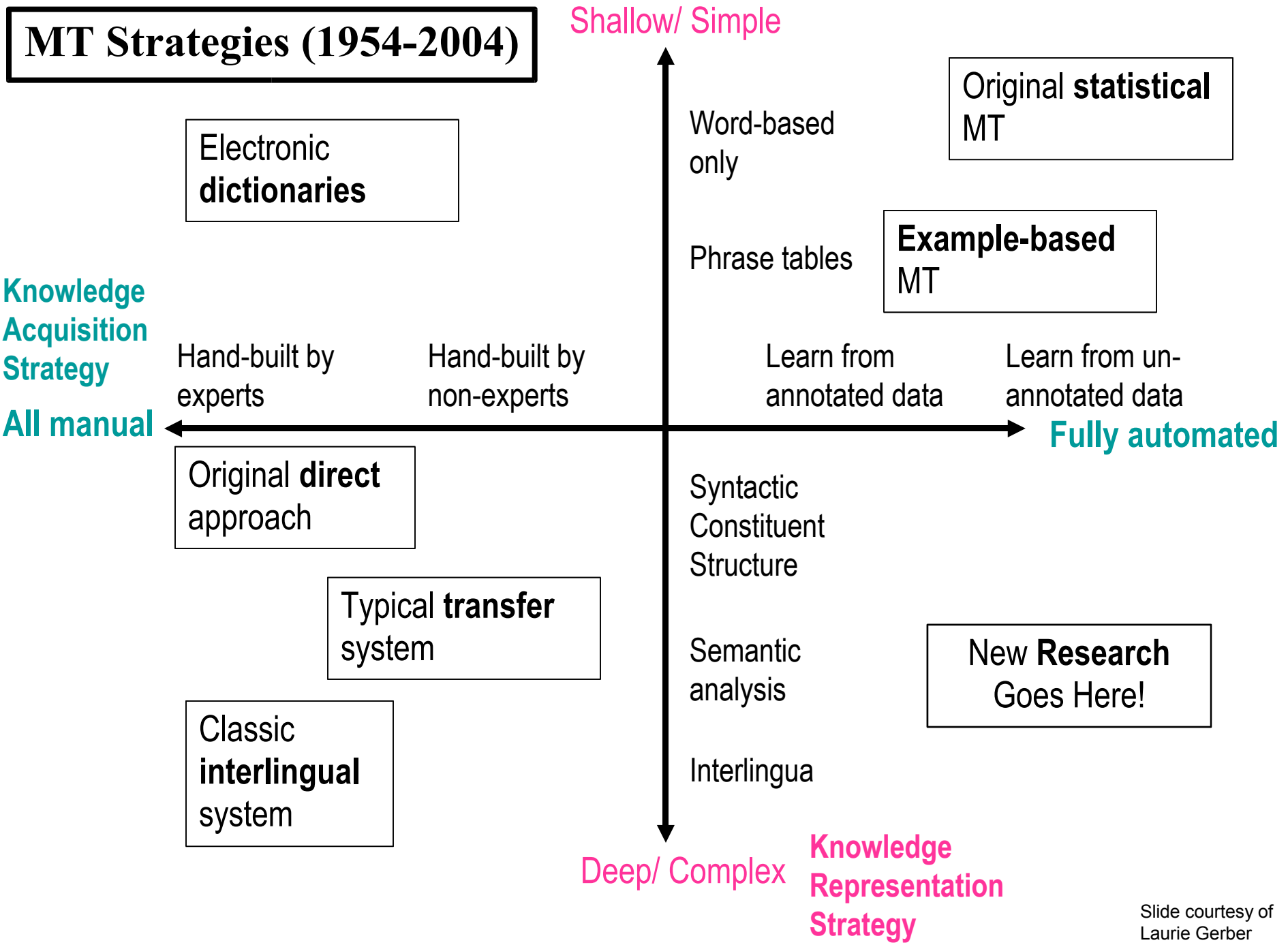
Cambridge MA, Aug 11, 2006

Hybrid Machine Translation Panel

Presenter: Nizar Habash (Columbia)

Moderators: Violetta Cavalli-Sforza & Alon Lavie (CMU)

Fellow Panelists: Jaime Carbonell (CMU), Philipp Koehn (Edinburgh), Stephanie Seneff (MIT), John White and Jean Senellart (Systran)



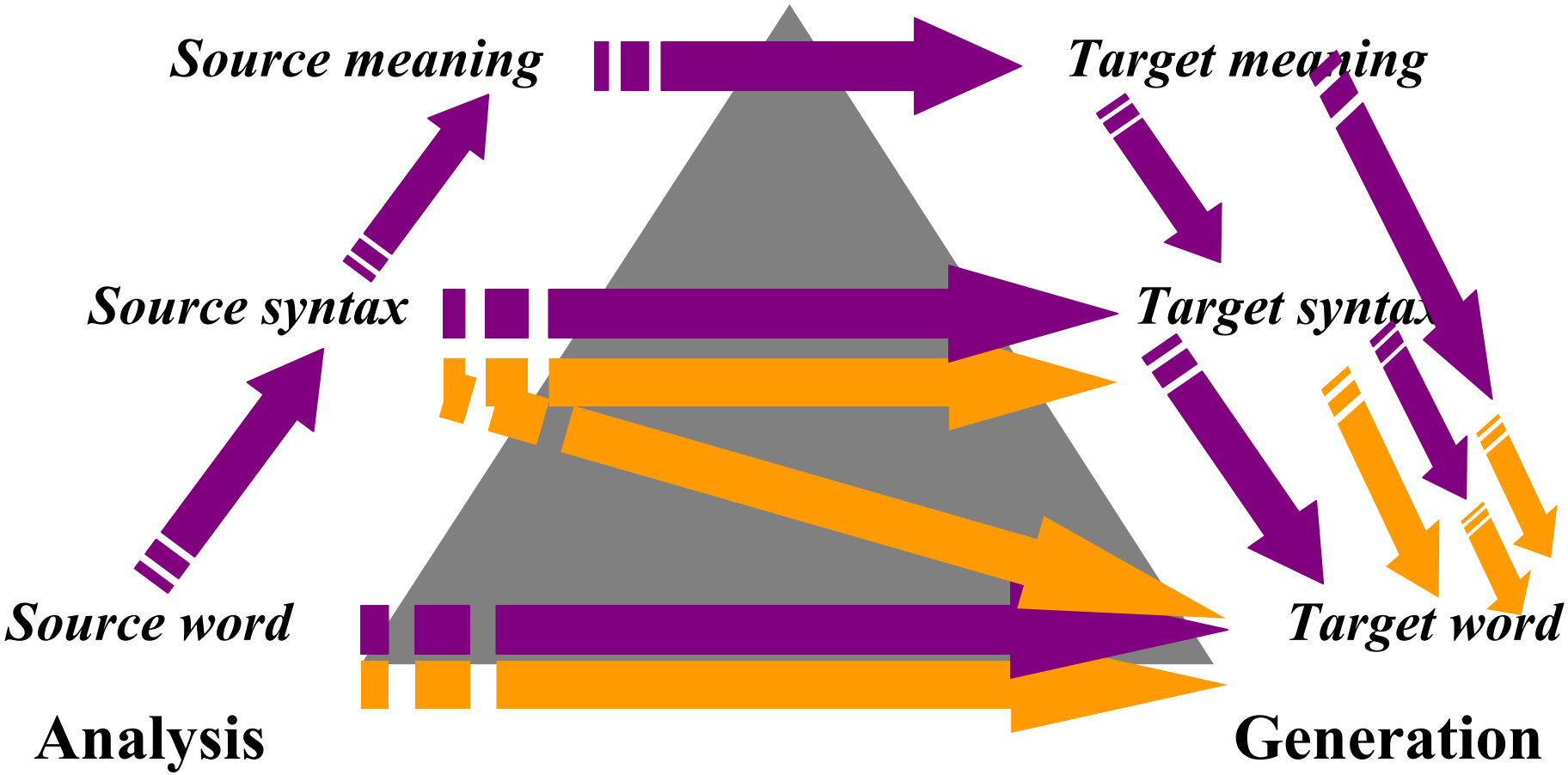
Slide courtesy of Laurie Gerber

What is a Hybrid MT system?

- "Hybrid" is a moving target
  - StatMT systems use some rule-based components
    - orthographic normalization, number/date translation, etc.
  - RuleMT systems nowadays use statistical n-gram language modeling
- Hybrid continuum
  - Different mixes of statistical/rule-based components
    - Richly annotated corpora created by linguists
    - Used for statistical POS tagging
      - » As part of a symbolic interlingual systems
      - » Preprocessing for phrase-based MT
    - With statistical language modeling component
  - Every component can be done in either approach
    - Typically developers use what is available

MT Approaches

Statistical vs. Symbolic vs. Hybrid



Why Hybridize?

- The Intuition
  - StatMT and RuleMT have complementary advantages
    - Syntactic structure produces better global target linguistic structure
    - Statistical phrase-based translation is more robust locally
- The Resource Challenge
  - Parallel corpora as models of ***performance*** vs. Dictionaries/analyzers as models of ***competence***
  - "More is better" is true for both approaches
    - Parallel corpora are domain/genre specific
    - Dictionaries and parsers can be domain/genre specific
  - Hybrids may need more data
    - Annotated resources

Why Hybridize?

- The Quality challenge
  - Current MT systems are not very good
  - Statistical MT Problems
    - Errors: *flew a plane carrying 241 passengers for four hours in indonesian without navigation systems*
    - Hallucinations: *A second Palestinian suicide bomber explodes himself in Baghdad*
      - "Palestinian" hallucinated
    - Limitations: *Unseen morphological forms unhandled; word-order limited by phrase size*
  - Rule-based MT Problems
    - Errors: *Ayatollah over/Ali Khamenei*
      - wrong POS → wrong parse
    - Hallucinations: *The storm totaled the crops*
      - 'totally destroyed' conflated semantically (Habash 2003)
    - Limitations: *Idioms/expression not in dictionary unhandled*

Hybridization Challenges

- Combination problems
  - Linguistic phrase versus statistical MT phrase

    ". on the other hand , the"
  - Meaningful probabilities for Linguistic resources (dictionaries/rules)
  - The linking up of different components with different representations
  - System complexity
- The best of two worlds?

  … the worst of two worlds?